Department of Economics

Psychological Factors and Labour Market Outcomes:

The Case of Immigrants and their Children in Germany

Anna-Elisabeth Thum

Thesis submitted for assessment with a view to obtaining the degree of
Doctor of Economics of the European University Institute

Florence, February 2011

EUROPEAN UNIVERSITY INSTITUTE
Department of Economics

Psychological Factors and Labour Market Outcomes: The Case of Immigrants and their Children in Germany

Anna-Elisabeth Thum

Thesis submitted for assessment with a view to obtaining the degree of
Doctor of Economics of the European University Institute

Examining Board:

Professor Richard Spady, Johns Hopkins University (External Supervisor)
Professor Jérome Adda, European University Institute
Professor Christian Belzil, Ecole Polytechnique
Professor Pedro Carneiro, University College London

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

Abstract

Europe is facing the challenge of integrating a growing number of immigrants and their offspring. On the one hand immigration can be a necessity to overcome problems arising from demographic changes but on the other hand cultural, social and political issues can hamper the economic potential immigrants have for their host society. In this introductory chapter I examine the bigger picture of immigration and integration in Europe and of the research on immigration. I show that there are several closely interlinked dimensions in the integration process and that - in terms of research on immigration - it is more interesting not to regard exclusively one dimension but to study one dimension in its context.

Abstract

Personality, ability, trust, motivation and beliefs determine outcomes in life and in particular those of economic nature such as finding a job or earnings. A problem with this type of determinants is that they are not immanently objectively quantifiable. They are rather concepts than exactly measurable and directly observable objects. There is no intrinsic scale - such as in the case of age, years of education or wages. Often we think of these concepts as complex and several items are needed to capture them. In the measurement sense, we dispose of a more or less noisy set of measures, which indirectly express and measure a concept of interest. This way of conceptualizing is used in latent variables modelling. I examine in this chapter in how far economic and econometric literature can contribute to specifying a framework of how to use latent variables in economic models. As a semiparametric identification strategy for models with endogeneous latent factors I propose to use existing work on identification in the presence of endogeneous variables and examine which additional assumptions are necessary to apply this strategy for models with latent variables. I discuss several estimation strategies and implement a Bayesian Markov Chain Monte Carlo (MCMC) algorithm.

Abstract

Educational attainment, length of stay, differences in national background and language skills play an acknowledged important role for the integration of immigrants. But integration is also a social process, which suggests that psychological factors are relevant. This chapter explores whether and to what extent immigrants and their children need to believe in their ability to control their own success, in other words their sense of control. To quantify this personal trait I use a measure of an individual's sense of control over outcomes in life - such as finding a job. This measure is known in

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

psychology as "the locus of control". I first estimate an exogenous measure. Then I address the problem that this measure is actually endogeneous in a labor market outcome equation by employing a model in which the sense of control is an endogenized latent factor in a simultaneous equation model. The determinants of this sense of control as well as its effect on the probability of being employed are examined. The model is estimated using an implemented Bayesian Markov Chain Monte Carlo algorithm. Results with endogenized personality indicate that, on average, immigrants believe less than natives in being able to control outcomes in life, but children of immigrants have already a stronger sense of control than their parents. The paper also finds that sense of control over life's outcomes positively contributes to the probability of being employed. This means that immigrants and their children face a double disadvantage on the labor market: they are disadvantaged because of their status as an immigrant and they have a lower sense of being able to control their situation, which is a personality trait that matters on the labour market.

Abstract

Identity can be an important driving force for educational performance. Immigrants and their children face the challenge of identifying with their host country's culture. This paper examines whether young immigrants and their children who identify stronger with the German culture are more likely to increase their educational outcomes. I use a concept of ethnic identity which is designed to capture Germanness in immigrants' day-to-day routine - based on self-identification, language skills and cultural habits. The research design takes into account the issue of endogeneity of ethnic identity in an educational outcome equation by measuring education and identity at different moments and by using an endogenous latent factor methodology. The paper finds that identification with the German culture has an overall positive effect on educational outcomes and diminuishes and renders insignificant the educational gap between immigrants and the second generation. The paper's results indicate that the second generation identifies stronger with the German culture than immigrants, no matter whether of German, European, Central European or Turkish background. Apart from the immigrant generation, own low educational attainment and high mother's educational attainment matter for identification with the German culture.

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

ACKNOWLEDGMENTS

I would like to thank my supervisor, Professor Richard H. Spady, sincerely. He has inspired me and given me the courage and motivation to explore new fields, to combine knowledge and expertise from different fields and to always go a step further than planned. He has also supported me very much on every level. I would like to thank my second supervisor Professor Jerome Adda and my former second supervisor Professor Pascal Courty for very helpful comments and questions. I would like to thank Professor Christian Belzil for very interesting discussions on latent variables and other topics. I would like to thank also Professor Andrew Chesher for an important discussion concerning the first chapter of my thesis. I would like to thank Marco Lombardi for his very helpful comments on the Markov Chain Monte Carlo methodology used in my thesis and Alicia Perez-Alonso for valuable discussions on identification. I would like to thank Professor James J. Heckman, who has also determined the path of my thesis by a very insightful conversation at the very beginning of my thesis. I would like to thank Thomas Liebig for making it possible to work for the OECD Division of International Migration during my PhD program. I express a special thank you to Thomas Bourke, Marcia Gastaldo, Jessica Spartaro, Julia Valerio and Lucia Vigna. I would also like to thank my friends at the EUI Economics Department, Mariya Teteryatnikova, David Scherrer, Konrad Smolinski and Michal Lewandowski for always being there. I would like to thank my parents and Arnaud Thysen for always supporting me.

CONTENTS

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

Part I

Introduction

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

> Because of differences in climate, energy, tastes and age, equality among people is a physical impossibility. But civilized man can render this inequality harmless, just as he has done with swamps and bears. (A. Checkov)

In the United States, the legend said, anyone could arrive and become a dishwasher and then even a millionaire. In Australia adventurous settlers left their maybe dark past behind and arrived on the fifth continent after a long journey on a boat and immediately found a patch of land to build a crop. After some fights against bush fires and wild animals these patches of land often grew to become prosperous and proud estates. These were the popular pictures of migration of about a century ago. Today immigrants need to fill in immigration cards, have a bunch of skills, pass tests, learn the language in classes and maybe even already have a job. The immigrant to Europe and also the intra-European immigrant does not face the wilderness of nature or the jungle of New York kitchens but he faces societies with little space, scarce offers of jobs, often mistrusting or even discriminating and imposing many integration rules. The difference is of course that the 18th century immigrant to Australia immigrated into a relatively 'new' country - sparsely populated inhabited mostly by aborigine clans - in which he could even define the rules, whereas the 21st century immigrant to Europe immigrates to settled societies. To be able to fight bush fires and to settle down successfully in a new environment or to find a job in a competitive and well established society requires psychological strength and cultural adaptation.

The aim of this piece of work is to study the role this psychological factor plays for the labour market integration of immigrants and their children. As the first chapter shows, the integration of immigrants is understood as a process involving several dimensions - a legal and political, an economic, a sociological, a psychological and a cultural dimension. It would be out of the scope of a dissertation in economics to study in depth the entirety of these dimensions and their interconnection. Nevertheless, I aim to integrate the psychological (and cultural psychological) with the economic dimension, taking the legal-political dimension as given. The economic dimension is understood as the one of interest, which is to be explained by the (cultural) psychological dimension. Throughout my dissertation I control for the different ethnic groups of immigrants in Germany. I consider two crucial indicators for immigrants' integration: employment and education.

To find a place in society, it is crucial for the immigrant to be employed. To find a job, beside socioeconomic determinants such as age and gender, the immigrant might need psychological skills. He might need to fight for an employment - so my hypothesis - by a high willingness to find a job and by a belief to find a job - a belief to be able to overcome obstacles which might destroy hope at the first instance. This relation between the economic target

of finding an employment and the psychological factor of believing in success is the topic of the third chapter. I find that immigrants and their children - controlling for socioeconomic differences - face a double disadvantage in society since they have lower levels of belief in their success than natives and this belief matters for economic success.

Human capital is another economic factor usually positively linked with employment. Education is also a transmission of values and it is a culture-specific element. Thus, young immigrants and the second generation, might find themselves confronted with a cultural conflict when following (part of their) education in a different culture. According to several sociological and psychological theories, this conflict might even hamper the educational attainment of individuals confronted with this conflict. The relation between their self-identification with the German ethno-cultural identity and educational attainment is studied in the fourth chapter. I find that the second generation identifies stronger with the German society than their parents and that identity is positively linked with educational attainment.

To study these questions, an econometric methodology is implemented, which has been developed recently by economics scholars, James Heckman and coauthors, on the one hand and by mathematics scholars, Ludwig Fahrmeir and Alexander Raach, on the other hand. Both methodologies are based on an estimator that has been constructed earlier in statistics to estimate a psychometric model. The methodology is interdisciplinary since it combines a psychometric with an econometric model. This approach is chosen to address the problem of measurement error that occurs when quantifying psychological traits and the problem of reverse causality between psychological traits and economic performance. In the second chapter, I study the place of this new methodology in econometrics and its possible added-value to economics and I propose a non-parametric generalization, using existing work on endogenous regressors.

In my dissertation, I could verify that psychological traits do play a role for the economic integration of immigrants and their children. In terms of employment, it helps to believe in being able to influence one's own success and immigrants and their children have on average a more fatalistic attitude in life. In terms of education, those immigrants and their children who adapt stronger with the German society are more likely to attain higher educational levels. I propose an econometric strategy - and critically analyze its added value, to incorporate other dimensions into an economic analysis in a statistically rigorous fashion. I understand my research as a contribution to a re-opening of economics towards insights from other fields, such as sociology, psychology and history, which seems to be one of the options of its development at the present time.

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

Part II

Thesis Chapters

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

CHAPTER 1

WHAT KIND OF IMMIGRANTS AND IMMIGRATION RESEARCH DOES EUROPE NEED?

In 2008 the Secretary-General of the OECD, Angél Gurria, has declared migration, health and water as the three priorities of the OECD during the first two years of his term[1]. According to the OECD (2007) four million new permanent immigrants entered the OECD countries in 2005 and in 2006 769 000 temporary migrants entered France, Germany, Italy and the United Kingdom taken together[2]. In 2009 there are 31 797 300 foreign nationals in the EU27 area which amounts to about 6.37%. The shares of the foreign-born population in the EU27 countries range from just below 5% in Hungary to above 35% in Luxemburg, as Figure 1.1. shows. These numbers do not include the children of immigrants, who were born in the host country but have a foreign background; so the numbers of individuals with a foreign background are even larger.

Most European countries have not implemented any selection processes for immigrants as for example New Zealand and Australia have. So an important question in Europe is what to do with the immigrants which are already present in the host countries and which are a growing part of many European societies. There is a range of suggested policies such as integration courses, language courses, policies to facilitate the labour market entry, antidiscrimination laws and affirmative action policies. In practice much has been done especially on the regional and community level [3]. On the other hand some countries such as France currently promote a policy to take away the French citizenship once an individual of foreign background has committed a crime. No unified European integration policy and political stance exist across Europe.

To determine which integration policies have the potential of being successful, it is not only important to define what the host country can offer immigrants to help them integrate but also to ask "How much the immigrant can contribute to his integration?". It is widely agreed upon that there must be integration policies designed by the European governments but the two-sidedness of the integration process should also be acknowledged. In the following

[1] See OECD (2007).
[2] See OECD (2009), Table II.5.
[3] See for example European Commission (2010).

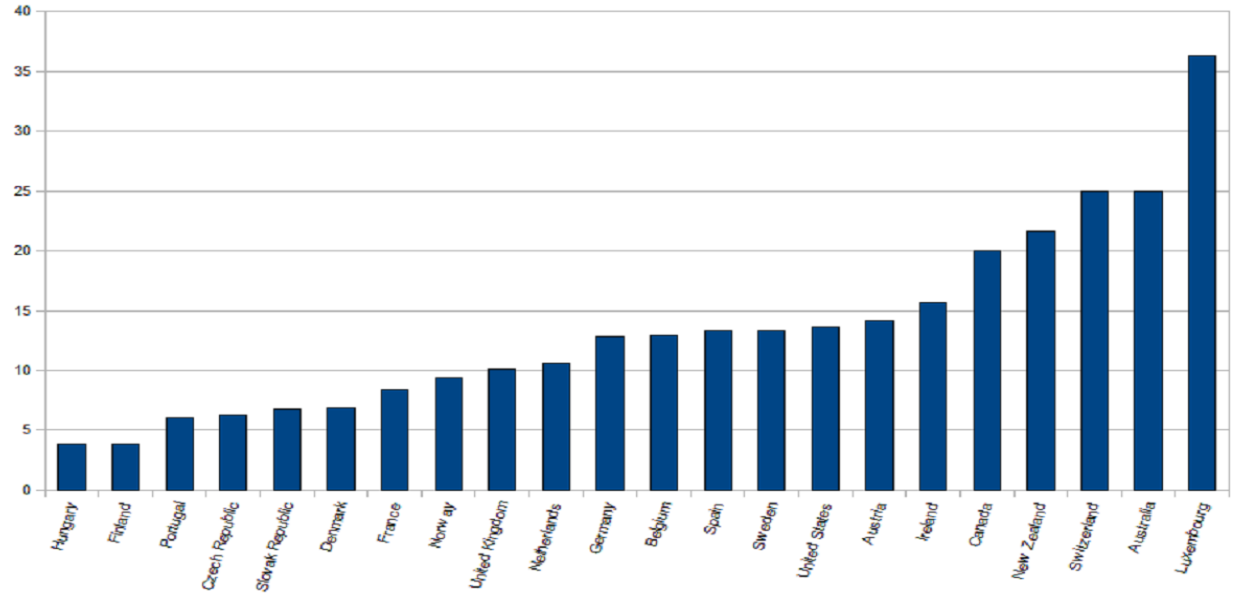

Figure 1.1: Shares of the Foreign-born Population in Percent of the Total Population in 2007, *Source: OECD International Migration Database*

chapters I ask the question, whether the immigrants (and their children) can also contribute to their success in the new society. In particular I set up and confirm the hypothesis that immigrants and their children are more likely to find a job once they believe more strongly in the possibility that their actions will help them to find a job. I also find that immigrants and their children, who identify more strongly with the host country culture will be more likely to further increase their educational levels. High educational levels are important because the probability of being employed is higher with higher levels of educational attainment, although even the highly educated immigrants and their children have not reached the employment levels of natives and their children in most countries[4].

Before going into the empirical and econometric methodological details in the following chapters, in this chapter I will explain why the immigrant needs to be an active part of the integration process and why research on the integration of immigrants needs to be interdisciplinary in the sense that a study of the economic factors needs to be linked to cultural and psychological factors.

[4]See for example Figure 1.3.

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

1.1 What kind of Immigrants does Europe need?

1.1.1 The Historical Origin of Europe's Immigrants

After the 2nd World War in 1945 there were large flows of workers especially into those European countries that needed to be reconstructed. Western European governments - for example Germany, France or Belgium - signed contracts with Southern European countries, which should send workers. These were the so-called "guest-workers" or "Gastarbeiter" in Germany. Belgium signed a contract with the Italian government in 1946 to send Italian labour to work in the mines and another contract with Morocco in 1964. In 1955 Germany signed a contract with Italy, in 1960 with Spain, in 1961 with Turkey and in 1968 with the former Yugoslavia. France also recruited workers from Southern Europe, but on a smaller scale. These workers were supposed to stay temporarily in the host countries, so no policies were designed to integrate them. The workers stayed however, which makes an integration policy necessary and which gave Europe a lesson on having failed to design proper integration policies at the time of recruitment as an article in the *Economist* puts it.

To Britain and France there were also the flows of immigrants from the former colonies in Asia, Africa and the Caribbean. In France, by 1970 there were 600 000 Algerians, 140 000 Moroccans and 90 000 Tunisians. About 300 000 immigrants came from the overseas colonies to France. Britain experienced large immigration from Eire and overseas immigration from the Caribbean, India, Pakistan, Hong Kong and Africa. Germany experienced a large inflow of ethnic Germans, who immigrated from the former "Reich" . After 1945 these were about 8 million people. They received German citizenship immediately since they were recognized as ethnic Germans. In the 1950ies Germany received about 3 million refugees from East Germany. In 2009[5] Germany had about 7.2 million immigrants - or about 8.76% of the population - of which 1.8 million were Turkish[6] and 150000 were from former Yugoslavia (compared to about 1 million in 2002).

Since 1945 immigration into Europe has become less Eurocentric and more culturally diversified. Questions of different dress codes enter even the political debates. Each country in the European Union has a different migration history and different conceptions of "nationality" and of "nation". This should be taken into account when trying to find a European policy of integration.

[5] See Eurostat, International Migration and Asylum Statistics.

[6] This number has decreased slightly since 1998 in which there were about 2.1 Turkish nationals in Germany.

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

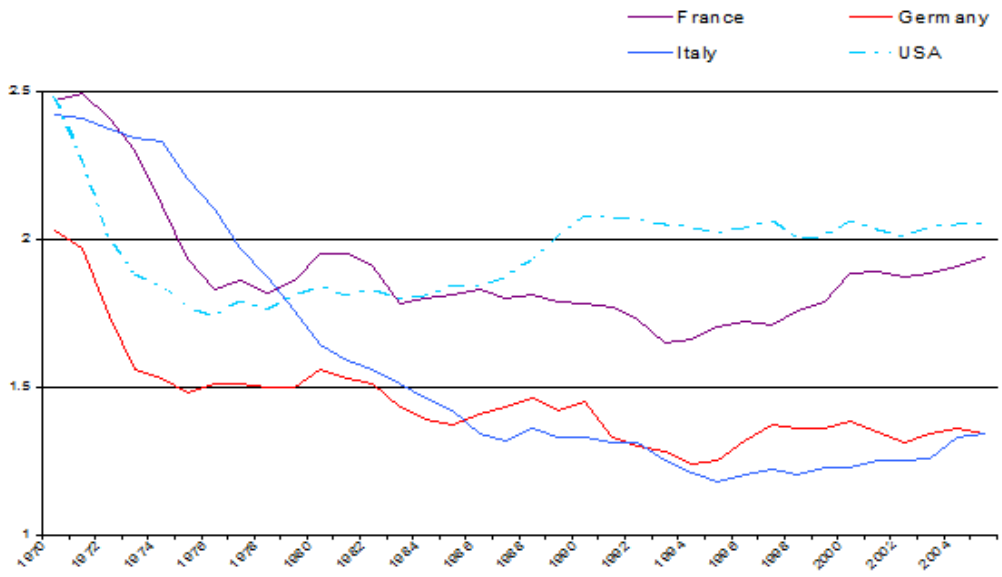

Figure 1.2: Fertility Rates in Selected Countries, *Source: Centre of European Policy Studies calculations based on OECD numbers*

1.1.2 Europe facing Immigration

Today, in the light of the demographic changes of European societies, immigrants play an important role for the future of European labour markets. As displayed in Figure 1.2 fertility rates in Germany have decreased from 2.1 in 1970 to 1.3 in 2006. In Italy the fertility rate was 2.4 in 1970 and has reached the same level as Germany in 2006. The consequence is an ageing European population. The European Commission (2008, p. 5) predicts that by the old age dependency ratio will rise from 22% in 2005 to 29% by 2020 and to 48% by 2050. One way to address these issues would be to change the paradigm of European societies and to introduce policies to encourage child bearing. A complementary and more rapid way to address the demographic problem is to encourage immigration.

It is widely acknowledged that immigrants should be employed and well educated in order to be integrated into the new societies. Being well educated is an advantage for employment. Numbers displayed in Figure 1.3. show however that even the most highly educated immigrant population has not reached the level of the native population. This phenomenon can have many reasons - discrimination, cultural adaptation and differences or difficulties in building a network for finding a job.

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

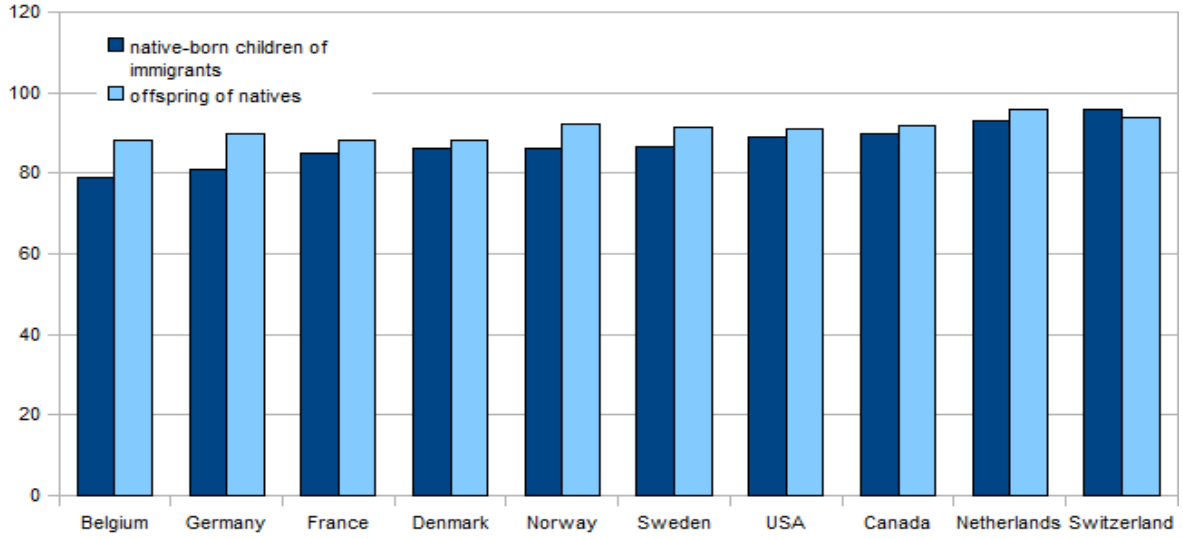

Figure 1.3: Employment Rates of Highly Educated 20-29 year old population in Selected Countries, around 2007, *Source: OECD*

In trying to explain the reasons for these numbers, it becomes clear that there might be a cultural and psychological dimension at play even for the economic integration of immigrants and their children. A considerable problem, which persists especially among the young generation of immigrants and children of immigrants is the role of individuals standing between two cultures. Problems such as riots in French suburbs may be attributed to difficulties in finding the way for young individuals with migration backgrounds. The current discussion in Germany which was launched by the former member of the board of directors of the German Central Bank (Deutsche Bundesbank) and former member of the Social Democratic Party Thilo Sarrazin's (2010) book "Deutschland schafft sich ab: wie wir unser Land aufs Spiel setzen" ("Germany does away with itself") on immigration and integration in Germany also shows that host country societies have not come to an end in terms of the cultural issues arising with growing numbers of immigrants in the host country societies.

1.1.3 Integration and the Immigrant - Psychological Paradigms of Integration

Facing these integration issues in the host country is the immigrant. He is a crucial part of the integration process. Integration is a two-way process. Rita Suessmuth, the politician, who was in charge of the 2000 immigration policy reform in Germany, wrote: "The

successful integration of immigrants ensures [..], that immigrants are able to participate in the economic, social, cultural, religious, political, and civic life of their host society and [...] that immigrants respect the fundamental norms and values of the host society and actively take part in the integration process." (see Suessmuth & Weidenfeld (2005))

Integration from the side of the immigrant involves psychological elements. One of these elements is "acculturation" - a concept this thesis examines in chapter 3 in its relationship to educational outcomes, which in turn are of concern for outcomes on the labour market.

Berry (1997) describes acculturation (a broader term for integration) as a dual process of "cultural and psychological change". The process involves for the individual to perform some smaller or larger change of behavioral aspects. Acculturation is not the same thing as assimilation as it was considered in the 1980s. A distinction is made between the acculturation of a society – as a response to confrontation with members of another society - and the psychological acculturation of an individual. Acculturation is related to the term interculturation, but less oriented towards creating a new culture.

Acculturation is seen as a multidimensional concept allowing the immigrant or immigrants' offspring to identify culturally with different cultures at the same time. On the individual level it is a highly variable process that can even lead to "cultural and behavioral loss in nondominant cultures". In this thesis I use a unidimensional concept in order to focus on the determinants of acculturation (measured by ethnic identity in chapter 3). Phinney describes this endogeneity of acculturation measures such as ethnic identity in his article "Acculturation is not an independent variable: Approaches to Studying Acculturation as a Complex Process"[7]. The determination of acculturation is given by changes in behavior, attitudes, feelings and beliefs. In both chapters 2 and 3 I aim to use measures taking these dimensions into account. Acculturation changes over time. Due to lack of time series data on the ethnic and psychological variables in my dataset, I cannot fully take this into account, but the econometric methodology aims at addressing the endogeneity arising from this issue. Phinney writes "Because acculturation is multifaceted and dynamic, it cannot be understood in any depth as long as we think of it as a single variable..Acculturation, like human development, is a process, not a variable." The aim of this thesis is to take account of this important insight within the framework of an economic and econometric model.

Another important element in the integration process are the children of immigrants and the parent-child transmission. The integration of immigrants' children depends on socioeconomic influences and on their cultural heritage and the parent-child relationship should incorporate both the parents' cultural traditional elements as well as the host society's fea-

[7] See Bornstein & Cote (2006).

tures.

In the following chapters I aim to capture the broad and vast sociological and psychological processes involved in labor market integration and in education. It is an attempt to propose an initial step for a formal and quantitative analysis of psychological and sociological elements in the economic integration of immigrants. The analysis is conducted in an econometric framework. The methodology addresses the main problem of endogenous variables and measurement error which arise in the framework of an econometric model. I aim to capture a whole process of initial elements, which determine the psychological and sociological variables, which then determine the economic outcomes.

1.2 What kind of Immigration Research does Europe need?

An interdisciplinary or multidisciplinary type of research. Integration takes place in different interlinked areas. As already outlined in the introduction, at the basis is the legal integration, reflected by the immigrants' work or residence permit. Linked to legal integration is economic integration, which is usually characterized by the employment status, wages and also by education since educational levels are highly linked with labour market success. An immigrant with a work permit will find it easier to find a job or the other way around. There is also political integration. In some countries immigrants may vote or immigrants achieve the right to vote once they have a residence permit. Being able to vote can increase a person's feeling of belonging or his knowledge about the society. Another measure of political integration may even be activity in local -or national in still rare cases -politics. Apart from legal, economic and political integration, there is the sociocultural integration. This dimension includes language competence, the contact to the media of the host country such as newspapers, friends in the host country and also ethnic feelings of belonging. Sociocultural integration can foster economic integration and the other way around. A fifth dimension of integration is psychological integration. It is measured by well being, self confidence and satisfaction. Clearly this dimension also has a link to the economic integration. Self confidence is found to be linked positively with employment in general. In turn, well being can be a result of having a job.

In this the following chapters I try to disentangle some of these dimensions of integration. I focus on the economic dimension, the integration into the labour market, but I attribute at the same time a determining role to education, sociocultural and psychological integration (described above). This theory is embedded in the political dimension of integration, which is why I examine it below. The legal dimension is taken as given.

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

1.2.1 Economic Paradigms of Integration

The economic dimension of integration is concentrated on the labour market - indicators such as wages and employment are mainly used to measure the economic integration of immigrants. It focusses on whether the immigrant contributes to the production processes of the host country and usually assumes the psychological dimension as either being proxied by the economic dimension or as being part of the error term. The cultural dimension is integrated in several current papers, which I will review below. I will review the literature on the labour market integration of immigrants and their children in more detail in chapter 3.2.

1.2.2 Political Paradigms of Integration

Our modern society is faced with the terms “pluralistic”, “multicultural”, “ethnic coexistence” and many others. Should immigrants be integrated or assimilated? Do we want to live in a “melting pot” or in a “salad bowl”? Do we integrate on a short term or on a long term basis? Do we integrate only on a pragmatic labour market oriented basis or also on a social/cultural basis? Who should be integrated?

In the western world there are two opposed integration models – that of traditional immigration countries such as Australia, Canada, the USA and New Zealand and that of new immigration countries in Europe. Among Europe the United Kingdom is a special case due to its large commonwealth area. One can also oppose the Anglo-American and the continental European model. The model of traditional immigration countries can usually be described as a "melting pot". It is a model of multiculturalism, in which all existing cultures blend into a new culture. This model might change to some extent in the United States in the light of the growing numbers of Hispanics.

The Anglo-American model is characterized by certain pragmatism, by positive action policies and by concrete antidiscrimination and equal opportunities policies. In the United Kingdom the “Race Relations Act” was ratified in 1976 and in the United States the Civil Rights Act. The attitude in the United Kingdom shifted away from concerns to a quest for institutional equality and for peaceful coexistence of different ethnic groups. It is relatively easy to gain access to the UK citizenship, but even non-naturalized immigrants seem to have better employment chances than in the rest of Europe. The British model can be described as a "salad bowl", in which different ethnic groups peacefully coexist.

The continental European model is not as clear cut – France follows a strategy of assimilation of the immigrants. The French model is often called the "Republican model" - all immigrants need to adjust to the French culture. Multiculturalism and different ethnicities

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

are not fully recognized notions in France. It must be said though that this paradigm of assimilation is becoming less radical. Italy and Austria seem more concerned about immigration and close their borders. The Netherlands have been the most open country towards immigrants beside the UK, but have also tightened their policies. (North)Western Europe, on the whole is more careful towards immigration, despite the economic and demographic necessities. Family reunion policy for example is used in the United States to attract skilled labour, whereas in Europe family reunion is seen as a process, which further floods society[8].

Germany, the country used as an example in this thesis, follows a more integrative approach, allowing for certain multiculturalism. On her most recent visit to Turkey, chancellor Merkel was discussing with the Turkish president Erdogan to start up Turkish schools in Germany. It was only in 2000 that Germany recognized that it is an immigration country and integration policies were starting to be designed. Nevertheless Germany is a European country with a considerable amount of migrants and their offspring. Now there are 2.3 million persons with a Turkish background living in Germany.

Resulting from the political paradigm, in Germany, as a continental European country, there is more mistrust towards immigrants than in traditional immigration countries. A belief of political weight is the belief among certain natives that immigrants take away their jobs is still persistent. One way for the immigrants and their offspring to overcome this mistrust can be sociocultural and psychological factors. In this thesis I aim to shed some light on this issue.

1.2.3 Sociological Paradigms of Integration

In sociology there are several theoretical paradigms for analyzing the integration of immigrants and their children. Each sociological paradigm provides a determinant for integration and may explain differences in integration levels. Penn & Lambert (2009) class these theories into seven different paradigms. The first paradigm is the theory of discrimination and prejudice. It is a classical sociological theory to explain the incorporation of minorities into a larger group. Discrimination is shown to be an important element in the Western world - Penn et al (1990) show that the economic situation of Asian immigrants to Britain was strongly influenced by discrimination. Discrimination can enforce the development of segregated ethnic enclaves.

This is closely linked to the theory of the "ethnic minority trap". Immigrants and their children might form enclaves and stay in a segregated part of the labour market, such as the low skilled sector. This might even lead to segmented assimilation. Immigrants define

[8] See Suessmuth & Weidenfeld (2005) page 9.

their identity by an identity oppositional to the majority. So if the majority is well educated an ethnic enclave could have an identity ideal of being badly educated. In this thesis I find a positive relation between a more German ethnic identity and education across both immigrant generations, which could be interpreted as some evidence of this theory. It means that feeling less German might are correlated with lower educational outcomes. The reason could be a feeling of opposition to the majority.

A third paradigm named by Penn and Lambert (2009) is ethnic asymmetry. There are differences between ethnic groups and they explain differences in integration. They also explain why different ethnic groups differ in a different way from the majority.

There are also differences in the nations incorporating the migrants. Each nation has a different way in approaching the issue of interethnic relations and each nation has a different degree of importance it attaches to nationality and to the notion of a "nation". This is an issue that needs to be taken into account, especially when studying the integration of immigrants across different European countries.

Social class is a paradigm predicting that disadvantages of being an immigrant are transmitted to across generations. This means that an intergenerational transmission of labour market success and educational attainment and integration may take place. In this thesis I study the first and the second generation and I find that the second generation does better in terms of the cultural and psychological integration and also on the educational and labor market level - adding the relevant controls.

Another sociological paradigm is that of gender - cultural differences in the gender roles transmit to the intergenerational outcomes. Especially in Muslim societies the role of women is still less labour market oriented than in Western societies. Nevertheless policy makers see an important chance in aiming at convincing immigrant women to integrate more into the host country society. This attitude is then hoped to be transmitted to their children and to thereby foster integration.

Cultural conflict is a relevant sociological theory for this thesis. Members of minorities may suffer of insecurities of their identity which has a negative effect on their integration performance. This thesis aims to examine this theory empirically. Is it true that immigrants suffer from less confidence than natives? Are there generational differences in confidence? Are there differences in ethnic identity across generations? What are the effects of confidence and ethnic identity for employment and education?

1.3 Conclusion

In this chapter I have examined the bigger picture of immigration and integration in Europe and of the research on immigration. I have shown that there are several closely interlinked dimensions in the integration process and that - in terms of research on immigration - it is more interesting not to regard exclusively one dimension but to study one dimension in its context. Immigrants are an active part of the integration process and this can improve their situation on the labour market. The next chapter will analyze which econometric method is suitable to study this interdisciplinary question and how this methodology fits into the current economic and econometric research and state of the art.

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

BIBLIOGRAPHY

[1] Berry, J.W. (1997) Immigration, Acculturation and Adaptation, *Applied Psychology.*

[2] Berry, J.W.; Phinney J.S. & Sam, D.L. (2006): Immigrant Youth in Cultural Transition – Acculturation, Identity and Adaptation across National Contexts, Lawrence Erlbaum Publishers London.

[3] Bornstein, M.H. & Cote, L.R. (2006): Acculturation and Parent-Child Relationships, Lawrence Erlbaum Publishers, London, UK.

[4] The Economist (2009): Immigration - Europe's dark past, November 10th 2009, *The Economist - European Politics, Charlemagne's Notebook.*

[5] European Commission (2008): Regions 2020 - Demographic Challenges for European Regions, Background Document to Commission Staff Working Document SEC (2008) 2868 Regions 2020, An Assessment of Future Challenges for EU Regions, *Directorate General for Regional Policy,* Brussels, Belgium.

[6] European Commission (2010): Handbook on Integration for policy-makers and practitioners, 3rd edition, *Directorate General for Justice, Freedom and Security,* Brussels, Belgium.

[7] OECD (2007): On the Move - International Migration. *DELSA Newsletter No 5, OECD,* Paris, France.

[8] OECD (2009): International Migration Outlook 2009, *OECD*, Paris, France.

[9] Penn, R. & Lambert, P. (2009): Children of International Migrants in Europe - Comparative Perspectives, Palgrave Macmillan, Hampshire, UK.

[10] Penn, R.; Martin A. & Scattergood (1990): The Dialectics of Ethnic Incorporation and Exclusion, *New Community.*

[11] Sarrazin, T. (2010): Deutschland schafft sich ab. Wie wir unser Land aufs Spiel setzen, *DVA.*

[12] Suessmuth, R. & Weidenfeld, W. (2005): Managing Integration: the European Union's Responsibilities towards Immigrants, Bertelsmann Stiftung, Guetersloh, Germany.

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

CHAPTER 2

PSYCHOLOGICAL FACTORS IN ECONOMETRIC MODELS: CONCEPTUAL AND METHODOLOGICAL FOUNDATIONS

2.1 Introduction

On an intuitive level it is clear that personality traits matter in life and for economic outcomes. A more self-confident candidate might outperform a candidate with a higher graduation grade or a candidate with higher level of motivation can acquire the job. But on a theoretical and empirical level the relation between personality and outcomes is not as clear-cut. Personality psychology studies personality and economics studies economic outcomes but research in studying the effects of personality on economic outcomes is still full of controversies and no consensual model has yet been found. There is a body literature in both economic theory and in econometric applications taking the relation between personality and economic outcomes into account and aiming at conceptualizing it and giving it an empirical back-up. Borghans et al (2008) give a detailed account on the potential of integrating insights and methodologies from personality psychology in economic and econometric models.

An example of a field of research in economics in which an integration of methods accounting for issues of social sciences, in particular the difficulty of quantification, as well as for issues of natural sciences, in particular the usually quantifiable nature of economic models, is immigration. As described in the first chapter, the integration of immigrants cannot really be reduced to a single dimension, as for instance the economic one. Immigrants face a new labour market which they need to understand and to which they need to adapt. This process needs to be taken into account for their economic integration. In particular, this piece of work shall examine the link between the psychological dimension and the economic one (in form of labour market outcomes) in the integration of immigrants.

In this first chapter I will examine the conceptual and methodological issues in studying the relationship between personality and economic outcomes, I will justify my choice of methodology with a view to the existing work, implement the estimation strategy used throughout the following chapters and assess identification possibilities of a generalized form of the model used in the following chapters.

I first outline the set up for a generalized form of the model to study the effect of

psychological concepts on economic outcomes. After discussing the interpretation of latent variables in econometric models and their added value, I present possible estimation methods of models involving latent variables. A special section is devoted to the Markov Chain Monte Carlo method to estimate parametric models including latent variable models, since this is the method I choose in the two applications in my thesis. I will then discuss identification possibilities and assess an existing semiparametric identification strategy for a model with endogeneity in its capacity to identify models with endogenous latent variables.

2.2 The Setting

To examine the effect of psychological concepts on directly measurable outcomes we propose the following model:

$$\begin{aligned} D_i &= \{0,1\} \\ D_i^* &= \beta^D X_i + \alpha^D \theta_i + \varepsilon_i^D \\ M_i &= \{1,2,3\} \\ M_i^* &= \alpha^M \theta_i + \varepsilon_i^M \\ \theta_i &= \gamma W_i + \varepsilon_i^\theta \end{aligned}$$

D_i is observable and signifies a discrete outcome such as the probability to be employed. It is modelled as a probit model with a latent underlying variable D_i^*. The outcome could just as well be continuous such as earnings. There is also a set of (for example tricategorical) observable categorical dependent variables M_i which signify the set of measures we use to measure the psychological concept. M_i is again modelled as an ordered probit model with a latent underlying variable M_i^*. Both D_i and M_i are affected by the psychological concept (equally called henceforth "latent variable" or "latent factor") θ_i. α^D and α^M express the effect of the psychological concept on the outcome and on its measurements respectively, henceforth called "factor loadings". In addition to θ_i there are observable explanatory variables X_i determining the dependent variable. Taking account of the the fact that psychological concepts are not exogenously given but are affected by social background and by experiences in life, there are observable variables W_i determining the latent variable θ_i. W_i and X_i can contain the same elements but cannot be exactly equal. If X and W are equal, identification cannot be guaranteed. One would need to add assumptions in this case. We discuss this issue in the next section since it is related to the assumptions imposed on the model.

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

2.2.1 Discussion of the Assumptions

In order to identify this model we need to impose several independence conditions, distributional assumptions and normalizations. In this section I will discuss and explain the general set of assumptions used throughout all following chapters. In section 2.5.3. we will relax some of these asssumptions. Throughout the two application sections - chapter three and chapter four - we will adopt a parametric strategy, but in section 2.5.3. we will examine the nonparametric case - for continuous and discrete outcomes - to take a more general view before applying the parametric case[1].

For simplicity in the following subsections I drop the subscripts noting the individuals.

2.2.1.1 Factor Analysis Assumptions

First of all, I will discuss the assumptions typically used in factor analysis. In the typical factor analysis with an exogenous factor, that does not depend on W as in our case, restrictions need to be imposed on the variance of the latent factor and on one of the loadings. This is to make sure that there is not an indefinite number of models that could be created. Typically it is assumed that $\theta \sim N(0,1)$; so the variance of the latent factor is assumed to equal unity. In our case θ depends on W and therefore the mean is not equal to 0 as is assumed in classical factor analysis, but it follows from the model that

$$E(\theta) = \gamma W$$

In the case of one latent factor as in the model above, we do not need to impose a restriction on one the loadings (Raach 2005: 21). However, we do impose that the variance of the latent factor is equal to 0:

$$var(\theta) = 1$$

2.2.1.2 Distributional Assumptions

We impose distributional assumptions on the error terms. The orthogonal random error terms $\varepsilon^D, \varepsilon^M, \varepsilon^\theta$ are all assumed to be follow $N(0,1)$. It follows from this assumption that θ is normally distributed:

[1] It would be an interesting further piece of research to implement the nonparametric case but would go beyond the scope of this thesis.

$$\begin{aligned} \varepsilon^D &= N(0,1) \\ \varepsilon^M &= N(0,1) \\ \varepsilon^\theta &= N(0,1) \\ \theta &\sim N(\gamma W, 1) \end{aligned}$$

2.2.1.3 Independence Conditions

A crucial point are the independence assumptions imposed. Raach (2005) and Fahrmeir and Raach (2006) do not mention these apart from a stochastic independence assumption for the θ equation (Raach 2005: 24). However Carneiro, Hansen and Heckman (2003: 377-383) provide a clear discussion for a similar model to the one above. Our independence assumptions are in line with Carneiro, Hansen and Heckman (2003) and include elements from Matzkin (2003).

For our model, first of all, we impose a standard assumption that the latent factor θ can be divided into an observable part γW and an orthogonal random part ε^θ with[2]

$$W \perp \varepsilon^\theta$$

Next we assume that

$$\theta \perp X|W$$

This assumption deserves some discussion. It is similar to the assumption (A-2) in Carneiro, Hansen and Heckman (2003:377). In their setup not the latent factor but the factor loadings depend on an a set of regressors, which they call state-specific regressors. Carneiro, Hansen and Heckman (2003) assume that the latent factor is independent of a vector of state specific regressors. In our setup the latent factor depends on a set of regressors, which makes it also similar to the setup in Matzkin (2003:6, 2007: 5326). Matzkin (2003,2007) assumes that there is a set of variables, which has the property that conditional on it an endogenous regressor becomes independent of an error part. We adopt this reasoning here by saying that there is a set of variables which we can include as determinants of θ, which control for the correlation between X and θ. From $\theta \perp X|W$ it follows that the correlation between X and θ is only through by W. To give an example, employment is dependent on education. Education is therefore part of X. A problem arises by the fact that a latent non-cognitive trait such as motivation might also depend on how educated an individual is. In this case X

[2] In the following exposition the symbol $\perp$ stands for "independent of".

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

and θ would be correlated and identification of the model would no longer be possible[3]. To overcome this problem, education is included both in the set X - determining the outcome D - and also in the set W - determining the latent trait θ. In this way the possible source of endogeneity of θ through a correlation with X is addressed.

$\theta \perp X|W$ is not an instrumental variable assumption, so W cannot be interpreted as an instrumental variable. An instrumental variable Z would fulfill the conditions that $cov(X, Z) \neq 0$ and $cov(Z, \varepsilon) = 0$.In our case W fulfills the conditions that $cov(W, X) \neq 0, cov(W, \theta) \neq 0$ and $cov(W, \varepsilon^M) \neq 0$.This latter condition is not in line with the instrumental variable assumption that $cov(Z, \varepsilon) = 0$.

In addition we assume the independence conditions

$$\theta \perp \varepsilon^M|W$$
$$\theta \perp \varepsilon^\theta|W$$
$$X \perp \varepsilon^D|W$$

From $\theta \perp \varepsilon^M|W$ and $\theta \perp X|W$ it follows that $\varepsilon^M \perp \varepsilon^\theta|W$. From $\theta \perp \varepsilon^D|W$ and $\theta \perp X|W$ it follows that $\varepsilon^M \perp \varepsilon^\theta|W$.

Finally we assume the local independence conditions (see for example Rabe-Hesketh 2004) which is typical for latent factor models

$$M \perp D|\theta$$
$$M_i \perp M_j|\theta \quad \forall i \neq j$$

From the latter - together with $\theta \perp \varepsilon^M|W$, $X \perp \varepsilon^M|W$ and $X \perp \varepsilon^D|W$ it follows that $\varepsilon^D \perp \varepsilon^M|\theta$.

2.2.1.4 Normalizations

Next, I discuss the normalization restrictions. In typical ordinal modeling it is standard, as Albert and Chib (1993: 671) impose, to restrict the variances of D_i^* and M_i^* to $v(D_i^*) = 1$ and $v(M_i^*) = 1$.In our case however, D_i^* and M_i^* depend not only on observable variables X but also on a latent variable θ, which is a random variable. So, following Raach (2005: 22) we do not standardize $v(D_i^*) = 1$ and $v(M_i^*) = 1$.We do however standardize some of the cutpoints, namely we impose that the cutpoints c of the J tricategorical items between

[3]Carneiro, Hansen, Heckman (2003:377) need to assume that the latent factor is independent of the covariates in the outcome equation.

category one and category two are equal to 0. This is necessary because in a traditional ordered probit model (assuming for the moment simplicity that there are no latent factors involved) under the normality assumption of the error terms in the ordinal model we can only identify the difference between the cutpoint and the intercept. Setting one cutpoint to 0 solves this problem (Raach 2005: 20).

$$c_{1j} = 0 \quad \forall j$$

These assumptions are adapted for each application accordingly below[4].

2.3 Interpretation of Latent Variables in Models of Economic Outcomes

The interpretation and correct use of latent variables in econometrics is not a clear issue. In econometrics, the notion of "latent variables" or "latent factors" does not yet have a clear position, even though they are already found in applications. In macroeconomics and in the financial literature latent factors are often used to capture unobservable factors influencing financial markets[5]. In micro-econometrics, there are several articles using them to capture unobservable skills. The work by Carneiro, Hansen and Heckman (2003) is a prominent example. Latent variables have a more established role in psychometrics where they were initially used to measure intelligence. The initial model was supposed to extract a measure of intelligence from a set of questions, which were usually verbal and arithmetic tasks. Later their use was extended to personality psychology. A main difference between psychologists and economists in this context is that the economist is interested in outcomes and the role that a personality trait can play for its determination. Borghans et al (2008)[6] show the problems of using the psychometric latent variable approach in econometrics. They also give credit to the work of Heckman et al for addressing some of the problems and somewhat adopting the latent variable approach to econometrics.

A common problem in econometric analysis is the fact that the econometrician can only observe a part of the factors relevant for an economic problem of interest - the problem of endogenous covariates. The famous example is the "ability bias" in the returns to schooling literature : if we cannot observe separately the effect of an individual's ability on his earnings, it will be captured by the measured effect of the education variable (for example years of education) and education will be endogenous in an earnings equation if one cannot control for ability. Bowles, Gintis, Osborne (2001) extend this argumentation from ability as a cognitive

[4] See the sections on assumptions in chapter three and chapter four.
[5] An example for the use of latent factors in macroeconomics is the work by Marco Lippi.
[6] Section III B of their paper gives account of the limits of the psychometric approach in economics.

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

skill to non-cognitive skills such as self-esteem or motivation. Bowles et al make it clear that even when controlling for ability in addition to conventional observable explanatory variables in an earnings regression there is still a large amount of relevant unobservable variation. They do this by calculating the size of variance unexplained by conventional observable factors and cognitive ability.

To resolve this problem we might seek a different variable,which can replace education, but is sufficiently correlated with it and not correlated with anything unobservable and relevant for the dependent variable. This would be a valid instrument. If we know several instruments for one endogenous variable, we can use a linear projection of the instruments on the endogenous variable to be replaced. The more abstract the perturbing unobservable concept is, such as self esteem, the more difficult it can become to argue that there is an instrument not correlated to it but correlated to education. It could be easier to just control for the perturbing concept even if it is unobservable. In the following I am interested in assessing the potential of "latent factors" - unobservable but measurable concepts that enter the economic model - to address the problem of endogeneity.

2.3.1 Latent Variables in Psychometrics

An overview from a psychometric point of view is given in Rabe-Hesketh and Skondral (2004). Generally, two strands of modelling settings with the presence of latent variables are taken : factor models and item response theory. DeLeeuw and Takane (1987) show that the models are equivalent in a one-dimensional parametric setting, assuming normality in a two parameter logistic item response theory model[7].

2.3.1.1 Factor Models

Structure Factor models assume the following structure underlying a matrix of items:

$$M_{ij} = \Lambda_j f_i + \varepsilon_{ij}$$

That is, the observable variables M are linear in latent factors f.The effect of f on M is captured by the factor loadings Λ_j.i is the observational unit and can signify for example individuals in micro-econometrics or points in time in stock market models. j is the indicator for the number of observable items.

Additionally it is assumed,that

[7]The two parameter logistic model in item response theory is explained in section 1.3.3.2.

- factors f and error terms ε are orthogonal to each other
- $\Lambda_j f_i \perp \varepsilon_{ij}$
- f_i and ε_{ij} are typically assumed to be standard normally distributed, so M_{ij} is normally distributed
- M_{ij} is implicitly assumed to be continuous but can be discrete. In that case a latent variable M_{ij}^* (not to be confused with f_i) is assumed. Suitable cutoff points need to be specified.

To determine, whether it is possible to fit the data according to the model, the correlation matrix of items needs to be analyzed.

Interpretation To understand how to interpret the model we look at the covariance matrix of M dropping the subscripts:

$$cov(M) = \Lambda' cov(f)\Lambda + \Sigma_\varepsilon$$

In the special case of two observable variables and one underlying factor we can write:

$$\begin{aligned} m_1 &= \alpha_1 f + \varepsilon_1 \\ m_2 &= \alpha_2 f + \varepsilon_2 \end{aligned}$$

and for the variance-covariance matrix

$$\begin{bmatrix} \sigma_{11}^m & \sigma_{12}^m \\ \sigma_{21}^m & \sigma_{22}^m \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} \sigma^f \begin{bmatrix} \alpha_1 & \alpha_2 \end{bmatrix} + \begin{bmatrix} \sigma_1^\varepsilon & 0 \\ 0 & \sigma_2^\varepsilon \end{bmatrix}$$

This follows from the assumptions of linearity and independence of factors and errors of and between each other.We can then write

$$\begin{aligned} \sigma_{11}^m &= \alpha_1^2 \sigma^f + \sigma_1^\varepsilon \\ & \quad etc \end{aligned}$$

Due to the standard normality of f and ε,we know that

$$\sigma_{11}^m = \alpha_1^2 + 1 \Longleftrightarrow \alpha_1^2 = \sigma_{11}^m - 1$$

So squared factor loadings show how much variance in the observable variable is explained by the latent factor.

To see another interpretation of the factor loadings, we can write

$$\begin{aligned} E(m_1\theta) &= E(\alpha_1 f + \varepsilon_1, f) \\ &= E(\alpha_1 f^2 + \varepsilon_1 f) \\ &= E(\alpha_1 f^2) \\ &= \alpha_1 E(f^2) \\ &= \alpha_1 \\ &= cov(m_1 f) \end{aligned}$$

That is, factor loadings can be interpreted as the covariance between the observable variables m and the latent variable f.Furthermore, if we standardize m, we can write

$$r_{m_1,f} = cov(m_1 f)/\sqrt{\sigma_{11}^m \sigma_1^\varepsilon} = cov(m_1 f) = \alpha_1$$

That is, factor loadings can be interpreted as the correlation between the observable variables m and the latent variable f.

Exploratory vs Confirmatory Factor Analysis Depending on how much structure the researcher is able to impose, factor analysis can be confirmatory or exploratory. In exploratory analysis no assumption is made on the number of factors explaining a set of items. The aim of exploratory factor analysis is to explain the total variance (unique and common variance) in the set of items by the smallest possible number of factors. In confirmatory factor analysis an assumption following from theory about the number of factors is made. The aim of confirmatory factor analysis is to explain the common variance among the set of items by a supposed number of factors.

2.3.1.2 Item Response Theory

Item Response Theory originates in educational psychology and can be seen as the first version of factor analysis with discrete dependent variables. The most common item response models are the Rasch Model,the 2pl and the 3pl model. They all have a similar specification and assume additivity and a logistically distributed error term. In the 2pl model the probability to answer "1" to an educational test is given by

$$\Pr(Y_{ij} = 1) = \frac{\exp^{\alpha_j - \beta_j \theta_i}}{1 + \exp^{\alpha_j - \beta_j \theta_i}}$$

where θ_i is the score an individual has on a latent ability scale - it is considered as a continuous latent variable. α_j can be interpreted as an item difficulty parameter and β_j as the discrimination parameter. A probit link in this model is also possible assuming normally distributed errors.

Usually item response models are estimated by conditional likelihood, marginal likelihood or conditional likelihood. These methods resemble maximum likelihood estimation, but involve an additional step of integrating out the unknown parameters θ_i. The method suffers from problems of joint consistency when letting the number of persons and of items become infinitely large (see Douglas (1997)).

Exploratory vs confirmatory Analysis As in factor analysis, the researcher can choose between assuming a number of factors or testing for an adequate number of scales. The latter is called Mokken scaling and is based on a concept of a total score of ordered items (see Mokken (1971)).

2.3.2 Latent Variables in Econometrics

There is an acknowledged concept in econometrics, which is close to the concept of latent variables. Tom Wansbeek (2000) shows a relation between the latent variable concept and the concept of measurement error, a concept that has already been examined in econometric theory. Matzkin (2007) and others further develop this relation between the concept of latent variables and that of measurement error. The virtue of this relation is that econometrically relevant results for the concept of measurement error can be used for the analysis of latent variables.

Measurement error can cause an estimation bias because the independent variables might be endogenous if the measurement error is correlated with the error term of the model.

Assumptions in economics are usually motivated either by theory or by (previous) empirical observations in a similar context as the model of interest. Assumptions on the nature of the latent variable cannot be based on the latter criterion since these variables can obviously not be observed and at least in economics there is not much experience with latent variables yet. So there are currently two views on how to make assumptions in the field of latent variable modelling. One is to argue, that the variable is latent and therefore we are relatively free in our assumptions. For instance, the support or the variance of the latent variable can be argued to be assumed freely. The second view is to require theoretical backing of the assumptions. This backing can come from other sciences, such as psychology, genetics or neuroscience. For example, we may argue that latent ability is genetic and therefore ex-

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

ogenous. To unify these two points of view one could argue that assumptions on the latent variable itself may be in a sense arbitrary[8], whereas assumptions on the relation to other variables should be based on theory.

Why? Suppose for a moment we see the latent variable simply as one specific part of the variation in the data[9]. What can we say about this variation? It is the variation in the data of interest the econometrician does not take account of by observable variables. Now we are interested in extracting the part of this variation which bears some informativeness in the sense either that we are interested in it or that it contains a relation to a variable in the model and we need to control for it to get unbiased estimates. The characteristic of a "variation " being informative comes from its being relevant for explaining a different "variation" - this relevance can be interpreted as relations (correlations) with other variables. So we are interested in extracting a part of the unobservable variation, which is correlated or uncorrelated with other variables in such a way, that we can interpret this part of the variation based on its correlations. In other words, assumptions on the relations of elements of the unobservable noise lead to its interpretability and should therefore be guided by theory. Note that the *nature* of the relation is again unknown and poorly theoretically founded. Therefore, even if the *existence* of the relation is arguably theoretically founded, the *nature* of the relation should be inferred from statistical relationships. A suitable approach used in econometrics is to assume a nonparametric relation. We thereby do not impose a possibly ad hoc parametric functional form on the relation between the latent variable and the observable variables. Assumptions on the nature of this element of the latent unobservable variation, which satisfies certain independence assumptions, are less relevant for its interpretation.

So, the interpretation of the latent variable, on one hand, is based on the independence assumptions. The other element for interpretation, as in psychometrics, are the items or psychometric questions and the strength of correlation between these and the latent concept. These are conventionally chosen on the basis of the criteria reliability and validity, in other words, whether the set of items reflects a latent concept and whether it reflects the concept of interest.

2.3.3 Problems

One problem is asymptotics : if we increase the number of individuals,we increase the number of parameters to be estimated (see Douglas 1997).

[8] Albert and Chib (1993) argue in a similar way to motivate setting the variance of a latent underlying variable in an ordered response model to 1.

[9] In a sense,we follow Matzkin (2003), who argues that "exogenous variation" can be used as an unobservable instrument in the presence of an endogenous covariate.

Another is that, as shown by Douglas (1997), the distribution of the estimated latent variable will never converge to its true distribution. This fact violates an assumption for most further analytical analysis, using the estimated latent variable as a fixed (in a way observed) element in a different model. This could be the case for instance if we aim to estimate the effect of latent ability on wages, having estimated latent ability in a separate model based on test scores.

2.4 Estimation in the Presence of Latent Variables

The presence of latent variables increases the amount of parameters to be estimated and the complexity of the likelihood function. Since the latent variable is unknown it is integrated out in the likelihood function. For a set of items (of which outcome can be seen as one item) $y = (y_1, ..., y_M)$ the likelihood function takes the form

$$\begin{aligned} p(y|\theta) &= \prod_{j=1}^{M} p(y_j|\theta_j) \\ p(y) &= \prod_{j=1}^{M} \int_{\theta} p(y_j|\theta_j) p(\theta_j) d\theta_j \end{aligned}$$

This integral needs to be solved numerically. There are several ways to do this.

2.4.1 Likelihood Approach

In the following section I will briefly mention two alternatives to MCMC to estimate the posterior density : the EM algorithm, which has been used much in the past to solve Bayesian models, before MCMC became popular, and quadrature, which is a deterministic technique to solve analytically intractable integrals. It does not rely on sampling techniques.

Sampling methods seem to be more powerful than quadrature, if the integral of concern is of higher dimensions, since sampling is independent of the number of function evaluations. This can be the case if a likelihood function conditional on more than one latent variable is of concern. A problem with Monte Carlo integration used for multidimensional integrals is that is biased and needs a large sample for the bias to decrease sufficiently.

2.4.1.1 EM algorithm

The EM algorithm specifies a rule, which implies alternating between computing the expectation of a likelihood function including latent variables as if they were observed (E-

step) and maximizing the expected likelihood from the E-step (M-step). The parameters from the M-step are then used in the next E-step. The algorithm is able to incorporate missing data and unobserved variables. There is no guarantee that the estimator converges to a maximum likelihood estimator. For multimodal likelihood functions the algorithm will converge to a local maximum. EM is partially Bayesian since it produces a point estimate of a latent variable together with a distribution of the latent variable.

Within sampling algorithms, MCMC seems superior to EM if the underlying model is more complex: the algorithm is likely to converge merely to local maxima if the likelihood function is multimodal.

2.4.1.2 Numerical Integration: Quadrature and Cubature Rules (Deterministic)

To approximate a complex function, the numerical value of definite integrals across the function can be calculated by an algorithm (combining evaluations of the integrand by a weighted sum). The collection of rules of this type are called quadrature for one-dimensional integrals and cubature for higher dimensional integrals. For now, let us write an approximation of $f(x)$ by numerical integration as

$$Q(f(x)) = \sum_{i=1}^{N} w(i) \int_{a(i)}^{b(i)} f(c)dc$$

where $w(i)$ are the weights assigned to each interval of integration and N denotes the number of intervals. So the numerical integration rule is characterized by the spacing of the subintervals and the number and weights of subintervals.

This procedure comes in hand if the integrand $f(x)$ is known only at certain points or if a formula for the integrand may be known. A small number of evaluation combined with a small error are desired for the numerical integration method. Gaussian quadrature is suitable if the function is smooth and the limits are well defined. In the following I will discuss different types of numerical integration.

2.4.1.3 Quadrature Rules based on interpolating functions

A function - typically a polynomial - is used to interpolate an integrand between point a and b. For a polynomial of order 0 an interpolating function passing through the point $((a+b)/2, f((a+b)/2)$ can look like this:

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

$$\int_a^b f(c)dc = (b-a)f((a+b)/2)$$

The polynomial can be of higher order. For more accuracy the interval can be divided in subintervals, which are approximated separately and added up (composited, iterated rule). Whether the subintervals are equally spaced on $b-a$, yields different sets of rules. (Gaussian quadrature is not equally spaced.)

2.4.2 MCMC

In this section I focus on the Markov Chain Monte Carlo methodology since it is employed for estimation of the two applications in following chapters of my thesis. With this method, the likelihood function is approximated by constructing a sample from it. It is a method used in Bayesian statistics. The Bayesian paradigm is a suitable environment to estimate models with latent variables since in Bayesian statistics latent variables are treated as random parameters to be estimated. Below I will first give a brief overview of Bayesian statistics. Then I will explain the MCMC algorithm and discuss its advantages and disadvantages.

2.4.2.1 Bayesian Statistics

This section gives a brief outline of Bayesian statistics. For further reading introductory textbooks on Bayesian statistics include Berry (1996) and Lee (2004)[10]. A comprehensive treatment of Bayesian statistics specifically in the latent variable context (and in psychometrics) is given in Rupp, Dey, Zumbo (2004).

In classic frequentist statistics the unknown parameters of a model are considered as unknown but fixed quantities. In the Bayesian paradigm however, the unknown parameters are considered as random variables, which follow a probability distribution. The aim of estimation in a Bayesian framework is therefore to estimate the probability distribution of the parameters, given the data.

Consider a set of parameters θ and a set of data y, then the probability distribution of the parameters given the data, and the main element of interest of the Bayesian statistician, is

$$p(\theta|y)$$

[10] See Raach(2005).

which is called the posterior distribution function. The posterior distribution function of a model is rewritten by applying Bayes' theorem

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta} \tag{2.1}$$

where the denominator is constant since it does not depend on θ.Therefore we can write

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

where $p(\theta)$ denotes the prior beliefs on the values of the parameters before the data is taken into account. A prior is an assumption on the probability distribution of a parameter formed before observing the data. It can be interpreted as a flexible assumption - since it expresses a belief the researcher has but can be revised if the data gives a stronger information about the parameter.

The posterior distribution is therefore proportional to the likelihood function $p(y|\theta)$ and the prior beliefs $p(\theta)$. The likelihood derives from the model specification. The prior beliefs need to be set by the researcher in such a way that the joint posterior is proper, which means that it should integrate to a finite value.

There are several possibilities of setting priors. If the researcher is uncertain, the prior can be set in such a way that it does not contain any information. This is called a flat or uninformative prior and could be for example the uniform distribution. In this case the posterior is proportional to the likelihood function and the analysis can be interpreted as a classical frequentist analysis. Priors can also be "subjective", which means that they derive from a theory. They can be "empirical" if they derive from data. If a prior is said to be conjugate, it is from the same class as the posterior. For example the normal distribution has this property.

When setting the priors the researcher needs to make a tread-off between strongly identifying the model by tightly set priors and to leave enough freedom for the data to give evidence on the model by leaving the priors loose enough. Due to the existence of priors for any parameter of the model and the possibility to tighten these priors, it has been indicated [11], that Bayesian methods can always be used to analyze a non-identified model. Nevertheless, the informativeness of the statistical analysis can become questionable if the priors are too tight and leave no possibility to infer information from the data.

Hypothesis testing in Bayesian statistics differs from frequentist analysis. Bayesian statisticians believe that their paradigm allows a more accurate way of testing for significance of

[11] See Poirier (1998).

the parameters. Once the parameter distributions are estimated, they can be characterized by their means, modes or variances. In addition, we can calculate the area of the distribution with 95 % of the probability mass - the central posterior density. This allows to make statements about the probability of a parameter lying within a certain region. Bayesians claim that these statements are more helpful than the frequentist confidence intervals[12].

I now turn to standard algorithms to calculate the posterior density described above.

2.4.2.2 Markov Chain Monte Carlo methods

Two events brought forward the use of Bayesian methods in statistics in the early 1990s, which had not been widely implemented due to the intractability of the posterior density and especially of the integral in the denominator in equation (1). It needed to be estimated by cumbersome techniques such as Gauss-Hermite quadrature or the Newton Raphson method. The increase in computer power together with a publication by Gefland and Smith (1990) on a computer intense but implementable Markov Chain Monte Carlo algorithm to calculate the posterior density made Bayesian methods, and especially MCMC, more and more popular in statistics.

MCMC, in contrast to the classical numerical optimization, is a simulation based technique and relies on random number generation since it solves the integral by sampling. Robert and Casella (2004) provide a thorough account of MCMC methods and Gilks, Richardson and Spiegelhalter (1996) show different possible applications of MCMC.

MCMC combines the two elements "Markov Chains" and "Monte Carlo integration", which I will outline below before I turn to explain one of the most prominent MCMC algorithms, the Gibbs sampler, which I implement in the following chapters.

Monte Carlo Integration Monte Carlo integration is a simulation-based method to solve an integral of the form

$$E_f[h(X)] = \int_{\chi} h(x)f(x)dx \tag{2.2}$$

where X is a random variable with probability distribution $f(x), \chi$ is the probability space and $h(x)$ is an arbitrary function of x. In the classic frequentist we are interested in point estimates of parameters but, as outlined above, in Bayesian statistics we are interested in estimating the posterior mean of a parameter θ_p.The formula for the posterior mean of a parameter θ_p takes a similar form as equation (2):

[12] See Raach (2005).

$$E(\theta_p) = \int \theta_p p(\theta|y) d\theta \tag{2.3}$$

where $h(X)$ is equivalent to θ_p and $f(x)$ is equivalent to $p(\theta|y)$.

To solve the integral in equation (2), Monte Carlo integration will provide an approximation of the integral by generating a sample $x^{(1)}...x^{(M)}$ from the distribution $f(x)$,evaluating the function h at each sampled value and calculating the average

$$\bar{h}_M = \frac{1}{M} \sum_{m=1}^{M} h(x^{(m)})$$

$\bar{h}_M$ converges almost surely to $E_f[h(X)]$ see Breiman (1992) in Raach (2005).

Equivalently, in our context, for calculating the posterior mean of θ_p by Monte Carlo integration, we need to compute the average

$$\bar{\theta}_p^M = \frac{1}{M} \sum_{m=1}^{M} \theta_p^{(m)} \tag{2.4}$$

where $\theta_p^{(m)}$ are randomly sampled from $p(\theta|y)$. This is not straight forward and we make use of the properties of Markov chains, which I will outline in the next section.

Markov Chains For a thorough account of the use of Markov chains in MCMC, see Robert, Casella (2004).

Markov chains represent random processes evolving over time. Consider a state space Ω and a random variable X_i holding different states in the state space. The Markov chain is a chain of realizations x_i of the random variables X_i. The change from one point in the state space x_i to the next x_{i+1} occurs with a certain probability. It can be seen that the chain represents probabilistic jumps through the state space from one state to the next. An important property of the Markov chain is that it has no memory of where it has been in the past. So it can be characterized fully by the transition kernel, which represents the probability to jump from one state x_i to the next x_{i+1} :

$$P(x_{i+1}) = P(X_{i+1}|x_i) \tag{2.5}$$

Equation (5) shows that the probability of jumping to a state x_{i+1} depends only on the previous state x_i.

As i goes to infinity and each jump of the chain occurs with the probability specified by the transition kernel, the Markov chain will reach a stationary distribution and the random

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

variables X_i will follow the stationary distribution of the Markov chain. Usually once we know the transition kernel of a Markov chain, the stationary distribution of the Markov chain follows from this. If the Markov chain fulfills essentially the irreducibility condition[13], this distribution π is stationary or invariant and satisfies

$$\pi(dx_{i+1}) = \int_{\Omega} p(x_i, dx_{i+1})\pi(x_i)dx_i$$

which states that if x_i is distributed according to the invariant distribution $\pi(x_i)$, x_{i+1} is also distributed like π.

In our setting, the state space is the probability space of the posterior density, which we aim to explore by creating a sample from this probability space. The idea underlying MCMC is to use Markov chains in the opposite way - not to first specify a transition kernel and derive a stationary distribution of the chain, but to first specify a stationary distribution and specify the transition kernel in such a way that this stationary distribution is obtained. The aim is to construct a transition kernel for which the stationary distribution of the Markov chain is equal to the posterior density of interest. The resulting Markov chain can then be interpreted as a sample from the posterior, as i goes to infinity.

Constructing a transition kernel of a Markov chain in this way produces a Markov chain that has a stationary distribution equal to the posterior density; this provides a sample $\theta_p^{(1)}...\theta_p^{(M)}$ from $p(\theta|y)$ which allows us to construct the average $\bar{\theta}_p^M = \frac{1}{M}\sum_{m=1}^{M}\theta_p^{(m)}$ in equation (4). Markov chain properties allow us to construct a sample and Monte Carlo integration is employed to take an average over this sample in order to approximate the joint posterior.

There are two prominent algorithms among MCMC methods, to sample from the posterior. The main challenge in constructing the algorithm is to specify the correct transition kernel such that the stationary distribution is equal to the posterior distribution of interest. One is the Metropolis-Hastings algorithm and the second is the Gibbs sampler, which is easier to implement[14] and is a special form of the Metropolis-Hastings algorithm. In the next session I will explain the implementation of the Gibbs sampler.

MCMC algorithm: the Gibbs sampler The Gibbs sampler (a special case of the Metropolis-Hastings (MH) algorithm) is an algorithm to generate random samples from a multivariate distribution. When the algorithm is used in the Bayesian context, this distribution is the posterior distribution. The MH algorithm draws random values from proposal densities and

[13] See Robert, Casella (2004) for more details.

[14] The advantage is that there is no need to adjust acceptance ratios for the drawn values before implementing the algortihm (see Raach(2005)).

accepts or rejects these according to the MH acceptance probability such that the detailed balance condition holds. If the acceptance probabilities are constructed correctly, the resulting sample is a Markov chain which has a stationary distribution equal to the target density.

For the Gibbs sampler the proposal densities are the full conditionals

$$p_{p|-p}(\theta_p|\theta_{-p})$$

for $\theta_1...\theta_p...\theta_P$ parameters.

Consider the target (posterior) density of a vector of parameters $p(\theta)$ and the parameters $\theta_1...\theta_p...\theta_P$ of interest. We begin with starting values $\theta_1^{(0)}..\theta_P^{(0)}$ and construct a Markov chain $\theta^{(1)}..\theta^{(M)}$ of length M. When the Markov chain has converged to its stationary distribution, the chain can be considered being distributed according to $p(\theta)$. This is the case because the transition kernels, so the proposal densities to draw the random values from, have been specified such that the invariant distribution of the resulting Markov chain is equal to the target distribution, the posterior density.

The Gibbs sampler algorithm is constructed in the following way:

1. choose starting values $\theta^{(0)} = (\theta_1^{(0)}..\theta_P^{(0)})$
2. repeat for $0, 1, ..., M$:
   draw $\theta_1^{(m+1)} = p_{1|-1}(\theta_1|\theta_2^{(m)}..., \theta_P^{(m)})$
   draw $\theta_2^{(m+1)} = p_{2|-2}(\theta_2|\theta_1^{(m)}, \theta_3^{(m)}..., \theta_P^{(m)})$
   :
   draw $\theta_p^{(m+1)} = p_{p|-p}(\theta_p|\theta_1^{(m)}, ...\theta_{p-1}^{(m)}, \theta_{p+1}^{(m)}..., \theta_P^{(m)})$
   :
   draw $\theta_P^{(m+1)} = p_{P|-P}(\theta_P|\theta_1^{(m)}, ..., \theta_{P-1}^{(m)})$
3. return $\{\bar{\theta}_1^{(M)}..\bar{\theta}_P^{(M)}\} = \frac{1}{M}\sum_{m=1}^{M}\{\theta_1^{(m)}..\theta_P^{(m)}\}$

The main challenge of the Gibbs sampler is to specify the transition kernels, or the full conditionals $p_{p|-p}(\theta_p|\theta_1^{(m)}, ...\theta_{p-1}^{(m)}, \theta_{p+1}^{(m)}..., \theta_P^{(m)})$ correctly. Below, I will show the implementation of a Gibbs sampler using an algorithm by Albert and Chib (1993), where the full conditionals are normal distributions.

But first I will discuss how to determine whether the Markov chain has converged and briefly some alternatives to the Gibbs sampler.

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

Convergence Diagnostics After the Markov chain has converged the random sample is considered to be drawn from the posterior distribution. To determine, whether the generated Markov chain has converged, there are several diagnostics. It is possible to diagnose non-convergence but convergence can never be proven. Any ergodic chain converges. An ergodic chain satisfies the property that any state can be reached from any other state in a certain number of steps. The speed of convergence depends on the form of the posterior, the smoother it is the faster the convergence. There is no rule for the number of iterations, sample size, number of parameters to guarantee convergence.

First of all one needs to look at the traceplots, which show the development of the draws for each parameter. If there is no trend in them and the draws reverse around the mode of the distribution, this is a first indicator of convergence. There are also more formal tools to assess the autocorrelation of the chain. Low or medium correlations are not a problem, but high autocorrelation can be an indicator that the chain has not converged. Cowles and Carlin (1996) give an overview over convergence diagnostics[15].

Convergence can be sped up by standardizing the variables[16], using a latent concept to summarize variables or using multivariate normal priors and by picking initial values close to the posterior modes. 100 000 iterations should be used for a model with a large number of parameters. Storage problems can be overcome by thinning the chain, that is storing only a fraction of the iterations.

2.4.3 Why MCMC?

Before turning to the implementation of the Gibbs algorithm in the next section I will briefly discuss why MCMC methods can be suitable in the latent variable context. First of all, a Bayesian treatment of latent variables is suitable, since latent variables can be considered in this framework as random parameters. They are random in the sense that they vary across individuals.

Secondly, as mentioned above (in section 3.4), asymptotic analysis is a problem in the presence of latent variables because of this increase in parameters to be estimated when the sample increases. Bayesian analysis does not rely as heavily on asymptotic results as classical frequentist analysis, since

Thirdly, if one is willing to make parametric assumptions, the Gibbs sampler is an easy to implement tool and requires less computation than numerical integration methods even though it also requires a high amount of computing time due to slow convergence, relatively

[15] See Raach (2005).
[16] The correlations between parameters are then easier to calculate.

to numerical integration.

A possible drawback is mentioned by Imbens (2009). The choice of the prior can be arbitrary. If there is much uncertainty about the parameters prior to considering the data, to address this problem, the priors should just be chosen to be flat or uninformative enough, in order not to give arbitrariness too much weight. Priors can also be seen as less restrictive than assumptions in the classic frequentist framework since they are flexible assumptions. If the data is more informative and gives other indications than the prior, it will overpower the prior in the posterior distribution.

Another problem mentioned by Imbens (2009) is that MCMC methods are need high computer power, which is less and less a problem due to the fast advances in computer power.

To give more reason to see the advantages of MCMC in the latent variable context, I will show an implementation of the Gibbs sampler.

2.4.4 An Implementation of the Gibbs sampler: Estimating an Endogenous Latent Variable Model

In the following section I will show some simulation results of a Gibbs sampler to solve a parametric model including latent variables. This implementation will be used for the two applications in the following chapters. It is strongly related to work by Albert and Chib (1993), Carneiro, Hansen and Heckman (2003), Heckman, Stixrud and Urzua (2006), Fahrmeir and Raach (2006) and Raach (2005).

The joint posterior distribution takes the following form:

$$\begin{aligned} & \prod_{i=1}^{N} f(\beta, \alpha, \gamma, \theta_i, M_i^*, D_i^*, c | M_i, D_i, X_i, W_i) \\ \propto \quad & f(\beta) f(\alpha) f(\gamma) f(c) \prod_{i=1}^{N} f(M_i, D_i, M_i^*, D_i^*, \theta_i | X_i, W_i, \alpha, \beta, \gamma, c) \end{aligned}$$

where $f(\beta)f(\alpha)f(\delta)f(\gamma)f(c)$ are the priors and the factor loadings and coefficients are written as $\alpha = (\alpha^M, \alpha^D)$ and $\beta = \beta^D$. M_i is a vector containing the polytomous psychometric items of the model, D_i is a scalar containing a binary economic outcome variable. The likelihood function can be simplified as

$$\begin{aligned}
&\prod_{i=1}^{N} f(M_i, D_i, M_i^*, D_i^*, \theta_i | X_i, W_i, \alpha, \beta, \gamma, c) \\
= &\prod_{i=1}^{N} f(M_i^*, D_i^*, \theta_i | X_i, W_i, \alpha, \beta, \gamma, c) \prod_{i=1}^{N} f(M_i, D_i | \theta_i, M_i^*, D_i^*, X_i, W_i, \alpha, \beta, \gamma, c) \\
= &\prod_{i=1}^{N} f(M_i^*, D_i^*, \theta_i | X_i, W_i, \alpha, \beta, \gamma, c) \prod_{i=1}^{N} f(M_i, D_i | c)
\end{aligned}$$

The first simplification follows from the application of the product rule. The second step follows from the fact that the ordinal responses M_i and D_i are determined solely by the underlying variables $M_i^*$and D_i^* and by the cutpoints c. The likelihood function can factored out into $f(M_i^*, \theta_i|.)f(D_i^*, \theta_i|.)$ since we made the conditional independence assumptions above. The factors of the likelihood function can be written as

$$\begin{aligned}
&\prod_{i=1}^{N} [f(M_i^*, \theta_i | \alpha, \gamma, c, M_i, W_i) \{ \sum_{k_M=1}^{K_M} 1(M_i = k_M) 1(c_{k_M-1} < M_i^* < c_{k_M}) \}] \\
&\prod_{i=1}^{N} [f(D_i^*, \theta_i | \alpha, \beta, \gamma, D_i, X_i, W_i) \{ \sum_{k_D=1}^{K_D} 1(D_i = k_D) 1(c_{k_D-1} < D_i^* < c_{k_D}) \}]
\end{aligned}$$

Each of the factors $f(M_i^*, \theta_i|.)$ and $f(D_i^*, \theta_i|.)$ needs to be multiplied by two indicators - and indicator which equals one if the observation $M_i(D_i)$ falls in category $k_M(k_D)$ and an operator indicating that $M_i^*(D_i^*)$ must fall between two cutpoints $c_{k_M-1}(c_{k_D-1})$ and $c_{k_M}(c_{k_D})$ according to its category.

θ is unobservable and will be estimated. To make the mechanism by which θ_i determines M_i^* and D_i^* perspicuous we integrate out θ_i and obtain the conditional distributions of M_i^* and D_i^* conditional on the parameters of the model and the data.

$$\begin{aligned}
f(M_i^* | \alpha, \gamma, c, M_i, W_i) &= \int_{\theta} f(M_i^* | \alpha, \gamma, c, \theta_i, M_i, X_i) f(\theta_i | W_i) d(\theta_i) \\
f(D_i^* | \alpha, \beta, \gamma, D_i, X_i, W_i) &= \int_{\theta} f(D_i^* | \alpha, \gamma, \beta, \theta_i, D_i, X_i) f(\theta_i | W_i) d(\theta_i)
\end{aligned}$$

As described above the Gibbs sampler is an algorithm which samples from the joint posterior distribution in a sequential way. The idea of the Gibbs sampler is to sample one of

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

the elements $M_i^*, D_i^*, \theta_i, \alpha, \beta, \gamma, c$ at a time, conditioning on the last sampled values for the remaining elements. This procedure is equivalent to sampling from a set of conditional distributions separately. Each conditional distribution is a posterior conditional distribution of a parameter given the last sampled parameter values and the data. These conditionals - each of them constitutes one step of the Gibbs sampling algorithm - are called "full conditionals". In the following, I will derive the full conditionals constituting the Gibbs sampler for the model of this paper.

2.4.4.1 The Posterior Conditional Distribution of the Latent Underlying Variables

Albert and Chib (1993) propose a data augmentation procedure to sample latent underlying variables in a threshold model. It follows from their work, that the full conditional for the latent underlying variable of the binary response is

$$\begin{aligned} f(D^*|\alpha^D, \beta^D, \theta, D, X) &\propto \\ \prod_{i=1}^{N} f(D_i^*|\beta^D X_i^D + \alpha^D \theta_i, 1)\{\sum_{k_D=1}^{1} 1(D_i^* &= k_D) 1(c_{k_D-1} < D_i^* < c_{k_D})\} \end{aligned}$$

$\alpha^D, \beta^D, \theta$ signify the last sampled values (or the initial values for the first iteration of the algorithm). It follows from the normality assumptions on θ and ε that $f(D^*|\alpha^D, \beta^D, \theta, D, X)$ is normally distributed with mean $\beta^D X_i^D + \alpha^D \theta_i$ and $V(D_i^*)$ normalized to unity. The latent underlying variable is distributed as the following truncated normal distributions

$$\begin{aligned} D_i^*|\alpha, \beta, \theta_i, D_i, X_i \tilde{} TN_{(-\infty,0)}(\beta^D X_i^D + \alpha^D \theta_i, 1) \text{ if } D_i &= 0 \\ D_i^*|\alpha, \beta, \theta_i, D_i, X_i \tilde{} TN_{(0,\infty)}(\beta^D X_i^D + \alpha^D \theta_i, 1) \text{ if } D_i &= 1 \end{aligned}$$

Similarly, the full conditionals for each the polytomous variables are

$$\begin{aligned} f(M^*|\alpha, \beta, \theta, c, M, X) &\propto \\ \prod_{i=1}^{N} f(M_i^*|\alpha^M \theta_i, 1)\{\sum_{k_M=1}^{K_M} 1(M_i^* &= k_M) 1(c_{k_M-1} < M_i^* < c_{k_M})\} \end{aligned}$$

The latent underlying variables of the polytomous indicators are distributed as the following truncated normal distribution:

$$M_i^*|\alpha, \theta, c, M, X \sim TN_{(c_{k_M-1}, c_{k_M})}(\alpha^M \theta_i + \beta^M X_i, 1)$$

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

2.4.4.2 The Posterior Conditional Distribution of the Factor Loadings

The full conditional for the factor loadings for D can be written as

$$f(\alpha^D|\beta,\theta,D,X,D^*) \propto f(\alpha^D)\prod_{i=1}^{N} f(D_i^*|\beta^D X_i^D + \alpha^D\theta_i, 1)$$

where we choose normal priors $f(\alpha^D) = N(0,1)$ and $f(\alpha^D) = N(0,1)$.If we rewrite the equation for D_i^* and M_i^*as

$$\begin{aligned} D_i^* - \beta^D X_i^D &= \alpha^D\theta_i + \varepsilon_i^D \\ M_i^* &= \alpha^M\theta_i + \varepsilon_i^M \end{aligned}$$

we can treat it as a normal regression model and derive for M_i and D_i

$$\begin{aligned} \alpha^M|\theta_i, M_i, X_i, M_i^* &\sim N\left[(\theta_i'\theta_i+1)^{-1}\theta_i'(M_i^* - \beta^M X_i^M), (\theta_i'\theta_i+1)^{-1}\right] \\ \alpha^D|\beta^D, \theta_i, D_i, X_i, D_i^* &\sim N\left[(\theta_i'\theta_i+1)^{-1}\theta_i'(D_i^* - \beta^D X_i^D), (\theta_i'\theta_i+1)^{-1}\right] \end{aligned}$$

2.4.4.3 The Posterior Conditional Distribution of the Direct Coefficients

Similarly to the procedure for the factor loadings, we can write the model as

$$D_i^* - \alpha^D\theta_i = \beta^D X_i^D + \varepsilon_i^D$$

For the coefficients, we choose to set diffuse priors as well. The full conditionals for the intercepts are, according to Albert and Chib (1993,p.671)

$$\beta^D|\alpha^D, \theta_i, D_i, X_i, D_i^* \sim N\left[(X_i'X_i)^{-1}X_i'(D_i^* - \alpha^D\theta_i^D), (X_i'X_i)^{-1}\right]$$

2.4.4.4 The Posterior Conditional Distribution of the Cutpoints

We assume a uniform prior for the cutpoints and can write for the full conditionals for the polytomous responses

$$c^M|\alpha^M, \theta, M, X, M^* \sim unif\begin{bmatrix} \max\{\max\{M_i^* : M_i = k_M\}, c_{M-1}\}, \\ \min\{\min\{M_i^* : M_i = k_{M+1}\}, c_{M+1}\} \end{bmatrix}$$

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

2.4.4.5 The Posterior Conditional Distribution of the Latent Factors

Similarly as for the procedure for coefficients and factor loadings, we can rewrite the model as

$$\begin{aligned} D_i^* - \beta^D X_i^D &= \alpha^D \theta_i + \varepsilon_i^D \\ M_i^* &= \alpha^M \theta_i + \varepsilon_i^M \end{aligned}$$

and treat it as a normal regression model,where θ_i is the parameter to be estimated. Carneiro, Hansen and Heckman (2003) specify a mixture of normals as prior for the latent factors. We treat the latent factors as endogenous depending on γW_i. We treat θ_i in the same way as M_i^* and D_i^* for which the priors are implicitly determined by the prior distributions of the other parameters of the model and by the assumptions on the distribution of ε_i^D and ε_i^M. The prior of θ_i is therefore implicitly determined by the priors on the other parameters of the model and by the assumptions on the distributions of ε_i^D, ε_i^M and ε_i^θ.

We can then derive the full conditional for the latent factor as:

$$\begin{aligned} & f(\theta|\beta, \alpha, c, \gamma, X, W, D^*, M^*) \\ \propto & \prod_{i=1}^{N} f(M_i^*|\alpha^M \theta_i, 1) f(D_i^*|\beta^D X_i^D + \alpha^D \theta_i, 1) \end{aligned}$$

We do not need to condition on M_i and D_i since they are implicitly known through M_i^* and D_i^* and c

$$\begin{aligned} & \theta|\beta, \alpha, \gamma, \delta, c, M, D, X, W, D^*, M^* \\ \sim & N \left[\begin{array}{c} \gamma W_i + (\alpha^{D\prime}(\alpha^D + \alpha^{M\prime}\alpha^M + 1)^{-1} \\ (\alpha^{M\prime}(M_i^* - \beta^M X_i^M - \alpha^{M\prime}\gamma W_i) + \alpha^D(D_i^* - \beta^D X_i^D - \alpha^D \gamma W_i)), \\ I - \alpha^{D\prime}(\alpha^{D\prime}\alpha^D + \alpha^{M\prime}\alpha^M + 1)^{-1}\alpha^D \\ -\alpha^{M'}(\alpha^{D\prime}\alpha^D + \alpha^{M\prime}\alpha^M + 1)^{-1}\alpha^M \end{array} \right] \end{aligned}$$

2.4.4.6 The Posterior Conditional Distribution of the Indirect Coefficients

The posterior we sample from can be written as

$$\begin{aligned} & f(\gamma|\theta, W) \\ \propto & f(\gamma) f(\theta|\gamma, W) \end{aligned}$$

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

The model for the latent variable is

$$\theta = \gamma W + \varepsilon_\theta$$

We assume a diffuse prior for the coefficient γ. Similar to the procedures above we get:

$$f(\gamma|\theta, W) \sim N((W'W)^{-1}W'\theta), (W'W)^{-1})$$

I simulated the data for $N = 1000$ and ran the algorithm for 100000 iterations. The table below in appendix B shows the results. The algorithm has converged since the traceplots of all estimated parameters do not show any trends. There are no evident identification problems since the posteriors are not flat, they all have a single mode and they do are not equal to the prior. The estimated values are always close to the true value and the standard errors show that the estimated values fall into a confidence interval around the true value.

2.5 Identification in the Presence of Latent Variables

Even if latent variables can be seen as an alternative to instrumental variable techniques, most approaches to identify models in the presence of latent variables rely nevertheless on the existence of an instrument (see Matzkin (2003,2007)). Carneiro,Hansen & Heckman (2003) provide a semi-parametric identification strategy of a simultaneous equation model with the presence of latent variables, which is not based on the existence of such variables.

There is a literature on nonparametric identification of models with endogenous regressors - models with measurement error. Latent terms might be considered in this literature but it is not the main interest to identify and estimate these terms and their effect on the observable terms.

2.5.1 Parametric Approaches

Identification in conventional parametric factor analysis uses the terms of the variance-covariance matrix of the observable variables and fits a linear and additive model to express this covariance matrix with a latent factor and a random error term. Additionally either the scale of the latent variables needs to be set or alternatively one of the factor loadings is set to a fixed term. Distributional assumptions are made on the distribution of the random error term and the latent variable. Rosenbaum (1984) establishes a condition for identification of parametric models involving latent variables, which says, that the number of parameters to be estimated needs to be equal to the number of covariances in the model.

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

In item response theory the assumption of conditional independence - independence between the items conditioning on the latent variable - yields identification. A parametric ordered response is assumed to underlie the observed response pattern. The distributional assumption on the error term in this model then yields the functional form of the probability of answering a specific category to a psychometric question.

Heckman, Stixrud, Urzua (2006) implement a version of the semi-parametrically identified model of Carneiro, Hansen, Heckman (2003). These authors embed a classic factor model into a linear model for economic outcomes. They use independence conditions, exclusion restrictions and distributional assumptions for the unobservable terms.

2.5.2 Nonparametric Approaches

In Psychometrics there is a nonparametric literature on identification and estimation of latent variable models. Pioneering work can be found in Rosenbaum (1984) and Holland, Rosenbaum (1986), Ramsay (1991) and Samejima (1979, 1981, 1984, 1988, 1990)[17] More recent work based on a total score of items is found in Molenaar & Sijtsma (2002).

In economics, Spady (2007) developed a strategy to infer a latent underlying scale, which is based on the notion of stochastic dominance, from a set of psychometric items concerning political attitudes. His method relies on a minimal set of assumptions. Matzkin (2003) develops nonparametric methods to identify functions for continuous and discrete dependent variables in the presence of endogenous observable explanatory variables and unobservable instruments. Endogeneity can result from omitted unobservable variables,measurement error or simultaneity. She claims it is a non-parametric version of the work of Heckman et al cited above[18].

2.5.3 Identification of the model in its generalized form

Latent variable modelling can express different conceptions. The latent variable in economics is most commonly a latent underlying variable governing an ordinal response. An interest in the effect of a latent variable on observable variables is fairly recent in economics. In other fields studying latent variables has so far been subject of mainly a parametric analysis. As mentioned above there is a nonparametric literature in item response theory, based on the total score of the items. In economics a total score is of less interest since economists are usually not interested in ordering the dependent variables by their degree of discrimination.

[17] References to the last two authors are given in Douglas (1997).
[18] Matzkin (2004) mentions this on page 3.

2.5.3.1 Continuous Outcome Variables

We are interested here explicitly in combining the existing literature in several fields to establish well-formulated conditions to identify semiparametrically the effect a latent variable has on observable variables. We are additionally interested in the interpretation of the latent variable, which we base upon the choice of dependent variables and on conditional independence assumptions.

In the following I explore, how the identification of the model with endogenous regressors in section two in Matzkin (2003) and section 4.1 in Matzkin (2007) changes when the endogenous regressor is considered as unobservable. We find that we can apply Matzkin's identification proof, but we need to add assumptions on the model for the unobservable regressor.

The model in its generalized form takes the form

$$\begin{aligned} Y &= g_1(\theta, \varepsilon_1) \\ \theta &= g_2(X^\theta, \varepsilon_2) \end{aligned}$$

θ is not independent of ε_1.Y is an observable continuous dependent variable, X^θ are continuous or discrete independent variables and θ is a continuous endogenous latent factor. $\varepsilon_1, \varepsilon_2$ are random error terms.

In the following we aim to identify the function g_1.

Assumptions In the following exposition the symbol $\perp$ stands for "independent of".

Condition 2.1 $\theta \perp \varepsilon_1 | X^\theta$ *(for first line in the proof below)*

In other words $F(\theta, \varepsilon_1 | X^\theta) = F(\theta | X^\theta) F(\varepsilon_1 | X^\theta)$.

Condition 2.2 *The function* $g_1(.,.)$ *is increasing in its second argument* ε_1*.(for third line in the proof below)*

Condition 2.3 *The conditional distribution* $F(Y|\theta = \tilde{\theta}, X^\theta = x^\theta)$ *is strictly increasing. (for invertibility of* $F(Y|\theta = \tilde{\theta}, X^\theta = x^\theta)$*)*

Condition 2.4 $F_{\varepsilon_1|X^\theta}(e_1) = U(0,1)$ *(normalization)*

Condition 2.5 $g_2(X^\theta, \varepsilon_2) = \bar{g}_2(X^\theta) + \varepsilon_2$ *(for identification of* $F(\theta | X^\theta = x^\theta)$*)*

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

Condition 2.6 $F(\varepsilon_2) = N(0,1)$ *(for identification of* $F(\theta|X^\theta = x^\theta)$*)*

From conditions 2.5 and 2.6 it follows that

$$F(\theta|X^\theta = x^\theta) = N(\bar{g}_2(x^\theta), 1)$$

Identification In the following we aim to identify the function g_1.

Theorem 2.7 *If conditions 1-3 are satisfied, then for all* X^θ, ε_1

$$g_1(\theta, \varepsilon_1) = F^{-1}_{Y|\theta,X^\theta}(F_{\varepsilon_1|X^\theta})$$

Proof.

$$\begin{aligned} F_{\varepsilon_1|X^\theta} &= \Pr(\varepsilon_1 \leq e_1|X^\theta = x^\theta) \\ &= \Pr(\varepsilon_1 \leq e_1|X^\theta = x^\theta, \theta = \tilde{\theta}) \\ &= \Pr(g_1(\theta, \varepsilon_1) \leq g_1(\tilde{\theta}, e_1)|X^\theta = x^\theta, \theta = \tilde{\theta}) \\ &= \Pr(Y \leq g_1(\tilde{\theta}, e_1)|X^\theta = x^\theta, \theta = \tilde{\theta}) \\ &= F_{Y|X^\theta=x^\theta,\theta=\tilde{\theta}}(g_1(\tilde{\theta}, e_1)) \end{aligned}$$

The second line follows from condition 1, the third line follows from condition 2. The fourth line follows from substituting $g_1(\theta, \varepsilon_1)$ by Y.Given condition 3 we can take the inverse of the last line and get

$$g_1(\tilde{\theta}, e_1) = F^{-1}_{Y|X^\theta=x^\theta,\theta=\tilde{\theta}}(F_{\varepsilon_1|X^\theta}(e_1))$$

Given the normalization $F_{\varepsilon_1|X^\theta}(e_1) = U(0,1)$ we get

$$g_1(\tilde{\theta}, e_1) = F^{-1}_{Y|X^\theta=x^\theta,\theta=\tilde{\theta}}(e_1)$$

■

This proof is in line with Matzkin (2003) and Matzkin (2007) with θ being unobservable in our case. Since θ is unobservable and we cannot condition on it the result is still incomplete.

We can then apply Bayes rule to eliminate the conditioning

$$F_{Y|X^\theta=x^\theta,\theta=\tilde{\theta}} = \frac{F_{\theta|Y,X^\theta} F_{Y|X^\theta}}{F_{\theta|X^\theta}}$$

We can then show that all three parts in the fraction can be identified. From the assumptions above, we have that

$$F(\theta|X^\theta = x^\theta) = N(\bar{g}_2(x^\theta), 1)$$

$F_{Y|X^\theta}$ can be estimated by any nonparametric estimator for conditional densities such as kernels. It remains to identify $F_{\theta|Y,X^\theta}$. In an item response model, similar to Spady (2006, 2007), $F_{\theta|Y,X^\theta}$ can be estimated via $p(M|Y,X) = \int p(M|Y,X)f(\theta|Y,X)d\theta$ when imposing the exclusion restriction that $p(M|\theta,X,Y) = p(M|\theta)$. By this estimation procedure one can recover $F_{\theta|Y,X^\theta}$.

In this section we have shown that by adding conditions such as condition 2.5 and 2.6 we can apply the identification results in Matzkin (2003,2007) to a model for an economic outcome with a latent endogenous factor when we dispose of a set of items measuring the latent factor.

2.5.3.2 Discrete Outcome Variables

In the previous section we have shown how one could use existing literature on nonparametric identification to identify the effect of an endogenous latent variable on a continuous outcome variable using cross-sectional data, such as earnings. Especially micro-econometric and typically psychometric outcome variables are often discrete, such as employment or answers to any qualitative question. It goes beyond the scope of this thesis to provide a new methodology to a nonparametric or semiparametric identification of the effects of latent variables on discrete outcomes, but I would like to point the reader towards existing literature in this field and show up several possibilities of approaching the problem.

Carneiro, Heckman and Hansen (2003) have developed a semiparametric identification strategy of factor models with discrete choices and continuous outcomes. They estimate the model parametrically, using an MCMC method. They assume that the latent factors are generated from a mixture of normal distributions. Error terms are assumed to be normal but they are theoretically nonparametrically identified.

Spady (2006, 2007) proposes yet another way of semiparametrically identifying and estimating a discrete choice model with latent factors. He uses discrete data on voting behavior and attitudes in the US and is able to estimate semiparametrically the effects of a cultural and an economic factor on US voting behavior. Spady specifies an item response theory model and imposes minimum assumptions on the distributions of responses as a function of the latent factor. His first assumption is responses of individuals with a higher position on the scale of the latent factor stochastically dominate the responses of those with a lower position on the scale of the latent factor. His second assumption is a monotonic scale rep-

resentation for the scale of the latent factor. His model can then be estimated by sieve maximum likelihood estimation.

In the field of psychology Douglas (1997) has also contributed to the theory of nonparametric identification and estimation of nonparametric item response models. He develops a methodology to simultaneously and nonparametrically estimate the latent factors and their effects on the responses. Douglas applies a kernel smoothing methodology to estimate the unknown quantities. This methodology has mainly been developed for ability testing framework in psychometrics.

We can also turn to the literature on non- or semiparametric identification of discrete choice models in the presence of endogenous regressors and see how one can possible reinterpret these models in order to make them fit into the framework of meaningful latent variables. Chesher (2007) has studied the issue of endogeneity and discrete outcomes. He shows, in a nonparametric framework, how to partially identify important structural effects with minimal assumptions. Lewbel (2000) proposes estimators for binary, ordered and multinomial response models, which can deal with endogeneity problems. His methodology builds upon one special regressor with a coefficient normalized to one and the existence of instruments. This methodology, however, relies heavily on the right choice of the special regressor.

The most suitable paper among the literature on nonparametric identification and estimation of discrete choice models with endogenous regressors seems as in the continuous case again Matzkin (2003). Matzkin (2003) develops in section 5.1 the discrete case of the continuous model, which we extended above. She proposes to use the same methodology as she uses for the continuous model to identify and estimate discrete choice models with an unobservable variable correlated with an observable variable. Matzkin refers to Blundell, Powell (2003) in this section. In the following I will summarize her approach, which makes use of a slightly modified version of Blundell, Powell (2003). They use a control function approach to estimate a single index binary response model and propose a setup of the following form

$$\begin{aligned} y_{1i} &= 1\{x_i\beta_0 + u_i > 0\} \\ &= 1\{z_i\beta_1 + y_{2i}\beta_2 + u_i > 0\} \\ y_{2i} &= E(y_{2i}|z_i) + v_i = \pi(z_i) + v_i \end{aligned}$$

where x_i and u_i are correlated and x_i is set of observable covariates and z_i is vector of instruments with

$$E(v_i|z_i) = 0$$

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

Blundell and Powell (2003) propose to use $\hat{v}_i$ to control for the endogeneity of x_i in the structural model. Their identifying assumption necessary for this step is

$$\begin{aligned} u_i|x_i, z_i &\sim u_i|x_i, v_i \\ &\sim u_i|v_i \end{aligned}$$

They also need to assume monotonicity and continuity of $F(x_i\beta, v_i)$ in its first argument. They can then show that they can identify and estimate the vector of coefficients β.

Matzkin (2003:15-17) uses a slight modification of this approach by using a variable $\tilde{Y}_2$ instead of $\hat{v}_i$ as control variable for endogeneity. She can show that the same estimation methods outlined in Blundell, Powell (2003) can be used when $\tilde{Y}_2$ is used instead of $\hat{v}_i$.

In the following I will follow the approach Matzkin (2003:15-17) and show how this approach can be used to identify a model with a latent endogenous variable. As in the case of a continuous outcome above - and as mentioned in section 2.2.1.1., I will not implement this semiparametric identification strategy in the applied chapters three and four but will show how a semiparametric identification strategy could be constructed using the existing literature on endogenous regressors.

$$\begin{aligned} M &= 1 \text{ if } g_1(X, Y, \theta) > \varepsilon \\ &\quad 0 \text{ otherwise} \\ M^* &= g_1(X, Y, \theta) \\ Y &= g_3(W, \eta) \end{aligned}$$

where

M - binary psychometric item

X, Y - observable explanatory variables (X is exogenous and Y is endogenous)

W - observable "control" variable

θ - latent characteristic (endogenous)

$\varepsilon, \delta, \eta$ - error terms

Next we need to impose some identification assumptions, which follow from Matzkin (2003: 15-17). The difference to her model is that θ is interpreted as a psychological latent trait, which is of interest for the researcher. Matzkin (2003) can show how to identify a function $\theta = g_2(W, \delta)$ determining this latent trait θ and how to use previous results to estimate $E(M = 1|X, Y, W)$ by using W as a control variable in the same way as Blundell and Powell (2003) use $\hat{v}_i$.

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

Assumptions

Condition 2.8 $\theta \perp (X, \varepsilon)$

Condition 2.9 $\varepsilon \perp (X, Y, M)$

Condition 2.10 *there exists a function* $g_2(.,.)$ *is strictly increasing in its last argument and a random error term* δ *such that* $\theta = g_2(W, \delta)$, *with* $\delta \sim U(0,1)$ *and* $F(\theta|W = w) = U(0,1)$ *at some value* w *of* W.

Condition 2.11 δ *is distributed independently of* (X, Y, W)

Condition 2.12 *one of the coefficients of* X *equals one*

Identification It follows clearly from this setup that the latent characteristic θ is correlated with the regressor Y, and θ is therefore endogenous. Matzkin shows in section 5.1 that under the first two conditions - under conditional independence assumptions between X and θ, one can identify $g_1(X, Y, \theta)$ as well as the distribution of (θ, Y).There are no restrictions that the function g_2 or on the distribution of ε or δ need to belong to parametric families.

Matzkin then shows that she can identify the model using a modified version of Blundell, Powell (2003) - as mentioned above, by substituting their control variable $\hat{v}_i$ by W. Given the assumptions above, Matzkin can rewrite the expression for $E(M = 1|X, Y, W) = G$ and estimate it using one of the estimators proposed in Blundell and Powell (20023). Once the estimators for G, β, γ are obtained, Matzkin shows that one can estimate the distribution of ε and the function $g_2(W, \delta)$.

I have shown, by reinterpreting and rewriting the model in section 5.1 in Matzkin (2003), that the latter can be used for the framework of latent endogenous variables in a discrete choice model. So far the model outlined above does not include a parameter α, which signifies the effect of the latent variable θ on the discrete outcome M. This would involve specifying some additional assumptions.

2.6 Conclusion

Latent variables are being used in some economic models, but there is not yet an established framework in economics of how to use them - how to interpret and identify them and which estimation strategy should be used. The econometric paradigm does not propose explicitly how to treat unobservable concepts, which are not straightforward to quantify. The unobservable is usually treated as an error term and there is no special interest in the

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

information included in this term. In this paper I explored, how the identification of the model with endogenous regressors in section 2 in Matzkin (2003) and section 4.1 in Matzkin (2007) changes when the endogenous regressor is considered as unobservable. I find that we can apply Matzkin's identification proof, but we need to add assumptions on the model for the unobservable regressor. I additionally proposed and implemented a Bayesian Markov Chain Monte Carlo estimator for an endogenous latent variable model and find satisfying results for estimated parameters of simulated data.

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

2.7 Appendix A: Tables

	loadings	true values	strd errors
m1	0.27	0.40	0.08
m2	0.34	0.20	0.27
m3	0.17	0.20	0.06
d	0.22	0.20	0.07

Table 2.1: Simulated Model: Loadings

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

	coefficients	true values	strd errors
m11	-6.07	-7.00	0.65
m12	0.24	0.30	0.04
m13	0.41	0.40	0.05
m14	0.15	0.20	0.04
m15	0.63	0.60	0.05
m16	-0.05	0.00	0.04
m21	-8.31	-7.00	1.40
m22	0.00	0.00	0.04
m23	0.27	0.20	0.06
m24	0.46	0.50	0.08
m25	0.15	0.10	0.05
m26	0.38	0.30	0.07
m31	-7.01	-8.00	0.60
m32	0.13	0.20	0.04
m33	0.08	0.10	0.04
m34	0.36	0.40	0.04
m35	0.03	0.10	0.04
m36	0.47	0.50	0.04

Table 2.2: Simulated Model: Direct Coefficients Tri-categorical Items

	coefficients	true values	strd errors
d1	-10.34	-10.00	0.81
d2	0.26	0.30	0.05
d3	0.44	0.40	0.05
d4	0.43	0.50	0.05
d5	0.22	0.20	0.05
d6	0.34	0.30	0.05

Table 2.3: Simulated Model: Direct Coefficients Binary Item

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

	coefficients	true values	strd errors
w1	0.07	0.00	0.09
w2	0.74	0.70	0.21

Table 2.4: Simulated Model: Indirect Coefficients

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

BIBLIOGRAPHY

[1] Albert, J.H. & Chib, S. (1993) : Bayesian Analysis of Binary and Polychotomous Response Data, *Journal of the American Statistical Association, 88 (422), 669-679.*

[2] Berry, D. A. (1996) : Statistics: A Bayesian Perspective. Duxbury, London, UK.

[3] Blundell, R. & Powell, C. (2003) : Endogeneity in Semiparametric Binary Response Models, *The Review of Economic Studies, 71 (3), 655-679.*

[4] Breiman, L. (1992) : Probability. SIAM, Philadelphia, USA.

[5] Borghans, L.; Duckworth, A.L.; Heckman, J. & terWeel, B. (2008) : The Economics of Psychology and Personality Traits, *Journal of Human Resources.*

[6] Bowles, S.; Gintis, H. & Osborne, M. (2001) : The determinants of earnings: A Behavioral Approach, *Journal of Economic Literature 39 (4), 1137-1176.*

[7] Canerio, P.; Hansen, K & Heckman J. (2003) : Estimating Distributions of Treatment Effects with an Application to the returns of Schooling and Measurement of the Effects of Uncertainty of College Choice, *International Economic Review 44(2), 361-442*

[8] Cowles, M. K. & Carlin, B.P. (1996) : Markov Chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association 91, 883-904.*

[9] Chesher, A. (2007) : Endogeneity and discrete outcomes. *cemmap Working Papers (CWP05/07).* Institute for Fiscal Studies, London, UK.

[10] DeLeeuw, J. & Takane,Y. (1987) : On the Relationship between Item Response Theory and Factor Analysis of Discretized Variables, *Psychometrika 52 (3).*

[11] Douglas, J. (1997) : Joint Consistency of Nonparametric Item Characteristic Curve and Ability Estimation, *Psychometrika 62 (1).*

[12] Fahrmeir, L. & Raach, A. (2006) : A Bayesian semiparametric latent variable model for mixed responses, *Psychometrika.*

[13] Gelfand, A. E. & Smith, A. F. M. (1990) : Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association 85, 398 - 409.*

[14] Gilks, W. R.; Richardson, S. & Spiegelhalter, D. J. (Eds.) : Markov Chain Monte Carlo in practice. Chapman and Hall.

[15] Heckman J.; Stixrud, J. & Urzua, S. (2006) : The Effects of Cognitive and Non-cognitive Abilities on Labor Market Outcomes and Social Behavior, *Journal of Labor Economics.*

[16] Holland, P.W. & Rosenbaum, P.R. (1986) : Conditional Association and Unidimensionality in Monotone Latent Variable Models, *Annals of Statistics 14 (4).*

[17] Imbens, G. (2009) : New Developments in Econometrics, Lecture 13, Bayesian Inference, *cemmap lectures, University College London.*

[18] Lee, P. M. (2004) : Bayesian statistics: an introduction (3rd edition). Arnold, London.

[19] Lewbel, A. (2000) : Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables, *Journal of Econometrics 97 (1), 145-177.*

[20] Matzkin, R. (2003) : Unobservable Instruments, *mimeo*, Northwestern University.

[21] Matzkin, R. (2007) : Nonparametric Identification, *Handbook of Econometrics Vol 6B.*

[22] Mokken, R. J. (1971) : A Theory and Procedure of Scale Analysis, Berlin, Germany: De Gruyter.

[23] Poirier, D. J. (1998) : Revising Beliefs in Non-identified Models, *Econometric Theory, 14, 483-509.*

[24] Molenaar, I.W. & Sijtsma, K. (2002) : Introduction to nonparametric item response theory, Sage Publications, Thousand Oaks, USA.

[25] Rabe-Hesketh, S. & Skondral, A. (2004) : Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models. Boca Raton, FL: Chapman & Hall/CRC.

[26] Raach, A. (2005) : A Bayesian semiparametric latent variable model for binary, ordinal and continuous response, Thesis Ludwig-Maximilians Universitaet Muenchen, Germany.

[27] Robert, C.P. & Casella, G. (2004) : Monte Carlo statistical methods (2nd edition). Springer, New York.

[28] Rosenbaum,P.R. (1984) : Testing Conditional Independence and Monotonicity Assumptions in Item Response Theory, *Psychometrika 49 (3).*

[29] Rupp, A. ; Dey, D. K. & Zumbo, B. D. (2004) : To Bayes or not to Bayes, from whether to when: applications of Bayesian methodology to modeling, *Structural Equation Modeling 11, 424-451.*

[30] Spady, R. (2006) : Identification and estimation of latent attitudes and their behavioral implications, *cemmap Working Papers (CWP12/06).* Institute for Fiscal Studies, London, UK.

[31] Spady, R. (2007) : Semiparametric Methods for the Measurement of Latent Atti-

tudes and the Estimation of their Behavioral Consequences, *cemmap Working Papers (CWP26/07)*. Institute for Fiscal Studies, London, UK.

[32] Wansbeek, T. (2000) : Measurement Error and Latent Variables in Econometrics, North Holland.

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

CHAPTER 3

LABOR MARKET INTEGRATION OF GERMAN IMMIGRANTS AND THEIR CHILDREN : DOES PERSONALITY MATTER?

3.1 Introduction

Does personality matter for integration of immigrants into the labor market? Intuition and common sense suggest a positive answer, but scientifically it is not as straightforward to tell, especially for the economist. This paper does not invent a new immigrant "homo oeconomicus" with a mathematically modelled personality, feelings, culture and other human features. Instead, it suggests a starting point to include an aspect of personality, a noncognitive skill, into an econometric model of labor market integration for immigrants. Measures of noncognitive skills and their inclusion into economic models have recently been studied in the economic literature - theoretically as well as empirically. This body of literature is still small but since it follows the trend of derationalizing the "homo oeconomicus", it is promising. Studies of the role of noncognitive skills have been undertaken for natives and there is a small literature on noncognitive skills for immigrants. The aim of this paper is to enrich this small body of literature by using an elaborated statistical tool to address problems of measurement error and endogeneity, which are very evident when working with measures of psychological concepts, such as noncognitive skills. This tool allows to endogenize measurement of noncognitive skills by taking into account its determinants. It additionally allows construction of a measure out of an informative set of measures of the noncognitive skill. This is an advantage since psychological concepts are more complex to measure than a naturally quantitative variable such as age or years of schooling.

The personality aspect, or noncognitive skill, we consider is called in psychological terms "locus of control", developed by Rotter (1966). We follow Heckman, Stixrud and Urzua (2006) in this approach. The locus of control is a measure for the degree to which an individual believes he (or she - it matters for both genders) has control over the happenings in his life[1]. It is represented as a scale reaching from "external" to "internal". A high *external* locus of control indicates that the individual believes that his life is controlled by

[1] Osborne-Groves (2006) has already used this measure. See Heckman, Stixrud and Urzua (2006).

forces outside of his own influence and he or she does not have a high feeling of controlling his life. A high *internal* locus of control however indicates that the individual believes strongly in his ability to control his life. One hypothesis of this paper is that an immigrant, who believes in controlling his outcomes has a higher incentive to provide the effort to integrate. The locus of control is strongly linked to the concept of "motivation": if individuals feel they have control over their lives, they believe in a causality between their actions and outcomes and this will motivate them to take actions.

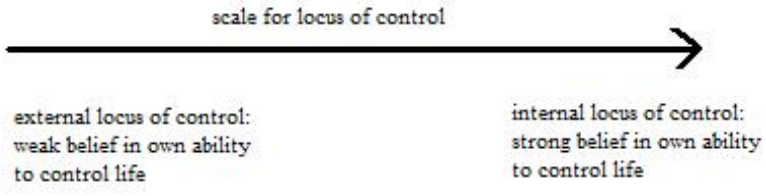

Figure 1: Measuring Personality - the Locus of Control

The locus of control of an individual develops over time, with foundations being laid in childhood and adolescence. Education, family background, maybe religion and also their personal immigration history - such as the arrival date in the host country - can play a role to determine an individual's locus of control. This is why we use a statistical tool, which allows to construct a measure of the locus of control, that depends on its determinants. In this way we address endogeneity problems of some variables, which could be included as controls in the employment equation and be correlated to the measure of the locus of control. This tool is based on work by Fahrmeir and Raach (2006).

This paper finds that a more *internal* locus of control has a positive effect on the probability of being employed. We find that being an immigrant has a significantly negative effect on having a more *internal* locus of control. Immigrants have on average a more *external* index of control. The same is true for the second generation, but the effect is not significant.

The paper is structured as follows. Section 2 gives an overview over the existing literature on labor market integration of immigrants and their children, and in particular in Germany, and over the use of psychological concepts in economics and in the literature on labor market

integration. In section 3 I introduce the data I use and describe the main variable definitions of the model. Section 4 presents the model. In section 5 I analyze the empirical results and section 6 contains conclusions.

3.2 Labor Market Integration of Immigrants and Their Children

The theoretical and empirical study of economic integration of immigrants was initiated by a seminal paper by Chiswick (1978), in which he shows that a catch-up process of "assimilation" of immigrants' earnings to those of the indigenous population takes place. This assimilation depends positively and crucially on the time spent in the host country. Among the first, Borjas (1985) shows that assimilation depends not only on the duration of stay but differs across cohorts. Borjas (1987) shows, that this cohort effect can be due to a change in the mix of countries-of-origin among the immigrant population.

A debate on how to study integration of immigrants economically commenced and a search for the main determinants of labour market integration began among mainly labour economists. A body of literature is concerned about different measures of integration, such as LaLonde and Topel (1992), Baker and Dwayne (1994) and Eckstein and Weiss (2002). Skills are considered as main factor of labour market integration, such as language acquisition in assimilation. A prominent example is work by Chiswick and Miller (1999). Human capital investment and transferability of skills as an important factor is, for example, studied by Duleep and Regrets (1999). Cultural Factors are studied by a smaller body of economic literature, but ethnic differences in immigrants' performance are acknowledged, for example by Chiswick (1988).

This literature mainly covers the United States and Australia. There is a recently developing body considering the German case, acknowledging the difference in labour market structure and immigration history with respect to the US. In particular, Dustmann and Schmitt (2000) study for example wage performance of immigrant women in Germany. Dustmann (1993) shows, that the status as a temporary vs permanent migrant has an effect on earnings adjustment. Constant, Zimmermann and Zimmermann (2006) examine the role ethnic self-identification. A comparative study of integration in Germany and Denmark by Hinte and Zimmermann (2005) provides a valuable overview over findings in the field.

Main findings on determinants of labour market integration in Germany are that formal education is particularly important since on the German labour market a high importance is attached to formal educational qualifications. Fertig and Schmidt (2001) argue that in the European context, discrimination might play a greater role for immigrants' earnings than in the US case.

In this research, in line with a body of research on the role played by personality as a determinant for earnings, I propose to investigate specifically on the importance of personality in form of non-cognitive traits for the labour market performance of immigrants.

3.2.1 Integration in Germany

In July 2006 the German Integration summit was held and was supposed to be the starting point for a national plan of an integration process. Preceding the summit was the ratification of a new integration policy, the Immigration Act of 2005. This is a nationwide policy program to integrate permanent migrants. These events provide evidence for the fact that German policy makers acknowledge the status of Germany as a country of positive net immigration.

Main goal of the national integration plan is to address language and educational deficits because of the high importance attached to formal degrees on the German labor market[2]. This is partly in line with policy recommendations of an OECD report[3] on the issue. However, the OECD recommendations stress the importance of vocational training, access to self-employment, improved organization of temporary work agencies and early language training. The OECD report also argues that anti-discrimination policies and initiatives should be introduced, especially to address the discrimination in terms of non-recognition of foreign degrees. According to Liebig (2007) there is no clear-cut evidence on discrimination in terms of wages. This picture could however be distorted by discrimination in terms of qualifications.

Germany has accepted immigrants since 1954. The rapid economic recovery after the war demanded a work force and contracts were made with those countries that provided workers – mainly Mediterranean. Even after the oil crisis in 1973, a large proportion of immigrants and their families, some of whom came under family re-unification policies, stayed on and became permanent residents. This phenomenon is a main origin for a second-generation of migrants, born in Germany of foreign parents. Between 1988 and 1993 another wave of immigrants, asylum seekers and ethnic Germans, came to Germany. The two waves of different types immigration add complexity in the assessment of integration if immigrants in Germany.

Up to beginning of the 1990s the labour market situation of the foreign born population was similar to that of the native born[4]. However from the beginning of the 1990s Germany

[2] See for example Gang and Zimmermann (1999).

[3] The report "The Labour Market Integration of Immigrants in Germany" (2007) by Thomas Liebig was published in the *OECD Social Employment and Migration Working Papers.*

[4] According to Liebig (2007) an exception was the employment rate of immigrant women and in particular

experienced a period of economic stagnation - with exception of the years 1998 and 2000 - and at the same time a particularly strong inflow of immigrants. The immigrants were hit strongest by the economic downturn[5]. Some steps to integrate immigrants were realized on a political level as early as 1974 when language classes for foreign workers were introduced. But the government nevertheless did not recognize Germany as an immigration country nor the necessity of the integration of immigrants up to the ratification of the Immigration Act mentioned above.

The developments would prove an integration policy useful. Elements of this policy would obviously include language promotion, equality of chances and social integration as measures of integration. These measures are supported if the immigrants have the will to integrate and to cooperate, in other words if they are motivated and believe in their success in integrating.

3.2.2 Labor Market Outcomes and Psychological Factors in Economics

The so-called bell-curve argument started by Herrnstein and Murray (1994) and their book "The Bell Curve: Intelligence and Class" in which they state, that IQ - or cognitive abilities - matters for socioeconomic success. They also argue that IQ is genetically inherited and is distributed across the population in form of a bell-curve and that some groups are marginalized in society due to their position within this curve. A response[6] to this discussion was given by Bowles, Gintis and Osborne (2001a, 2001b), henceforth BGO. In their work BGO claim that not only the cognitive skills matter but also personal traits and subsequently they provide a theoretical foundation and empirical evidence for their statement.

BGO believe that there are four puzzles in the wage determination literature which can be solved by taking into account personality traits as determinants of earnings. Firstly, earnings differ among apparently similar individuals. According to BGO in the United States only up to a third of the variation of log earnings are explained by the classical earnings determinants age, gender, education, parental education and occupation. Secondly, the advantages that successful parents transmit to their children go beyond the possibility to offer higher educational quality and the genetic transmission of cognitive skills. According to BGO studies seeking to explain the covariance of parental economic status and the respondents' income find considerable unexplained variance even after taking into account measures of IQ and the quality of schooling. Thirdly, there are supposedly irrelevant personality traits

of Turkish immigrant women.

[5] According to Liebig (2007) immigrants experienced a decrease in employment rates by ten percentage points, whereas the natives' employment rates decreased only by three percentage points.

[6] Another prominent response was "Lessons from the Bell Curve" by Heckman in 1995.

(such as home cleanliness) that matter for wage determination. Apparently, variables such as height, beauty or obesity are often robust determinants in earnings equations. Fourthly, findings on the effectiveness of school resources are controversial. Some authors find a positive impact on later earnings but not on educational performance. BGO provide evidence that non-cognitive traits might explain these puzzles. For example, managers rank such traits high on a list of desirable attributes a worker should have. They also cite a series of papers by Heckman and coauthors[7] which study the labor market success of General Educational Development (GED) diplomas. GED diplomas are tests are usually taken by high-school drop outs. The authors find that the cognitive skills of these drop-out are higher than those of high school graduates, but that GED graduates lack non-cognitive skills and usually have behavioral difficulties.

There are many personality traits and it is a complex task to determine those most relevant for labor market returns. BGO summarize this literature to some extent and find, that the traits that matter, are motivation as opposed to fatalism - measured in this paper by the locus of control, communication skills, attitudes, self-esteem and conscientiousness. A preference for challenge and fear of failure have an effect and can be seen as the risk attitude of an individual. Borghans, Duckworth, Heckman and terWeel (2008) have written an exhaustive account of the connection between personality psychology and economics and examine the explanatory power and interpretation of personality traits in economic models. They extensively review the relevant literature. The Five -Factor model on the five major dimensions of personality can also be an indicator of which noncognitive skills or personality traits to use. This literature claims that the five traits of extraversion, agreeableness, conscientiousness, emotional stability and openness constitute a personality. Farkas (2003) argues that additionally to the five factors, leadership, sociability and social sensitivity have an effect on socioeconomic success of individuals. Mueller and Plug (2006) estimate the effect of all five elements of the five-factor personality theory[8] on earnings of Wisconsin high school graduates. They find that all five elements have a significant effect on earnings. Osborne (2000) studies the impact of personality on labor market outcomes, their intergenerational transmission and differences in returns to personality across gender and occupations.

Motivation, measured in this paper by the locus of control, is a trait studied widely

[7]The authors name Heckman, Hsee and Rubinstein (1999) "The GEDas a Mixed Signal", *University of Chicago* and Cameron and Heckman (1993) "The Non-equivalence of the High School Equivalents", *Journal of Labour Economics.*

[8]The five factors are extroversion, agreeableness, conscientiousness, neuroticism, and openness to experience.

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

among researchers on the relationship between labor market success and non-cognitive traits. Duncan and Dunifon (1998) for instance have written a study of the long-run effects of motivation on labor market success.

The literature seems to agree that social background in turn plays a role in determining the personal traits affecting socioeconomic success and that noncognitive skills develop over time Cunha, Lochner and Masterov (2006) and Cunha and Heckman (2007) and Cunha, Heckman and Schennach (2010) develop a framework to model the change in noncognitive skill over time. Cunha, Heckman and Schennach (2010) model the development of both cognitive and non-cognitive skills as production functions at different time periods. Inputs into this production process are the parental environment, investments made at each period and initial personal endowments. The production process is modelled by a dynamic factor model with endogenous inputs. An advantage of their methodology is that they do not rely on test scores, which have no natural scale. They identify the scale by estimating the effect of the latent factors on adult outcomes. The authors find that interventions are more successful at the early stages of cognitive development, whereas noncognitive skills can also be successfully formed at later stages.

In the literature on immigration and integration of immigrants, and especially in the psychological literature, acculturation and identity strategies are seen as key factors for integration (see Berry (2001)). Sociability is also studied as a key factor for integration by dePalo, Faini and Venturini (2006). Fertig (2004) analyses the differences in leisure-time activities and attitudes of foreign immigrants, ethnic Germans and different generations and finds that both generations have differences in attitudes compared to Germans. Second generation immigrants seem to be the most fatalist and pessimist.

This paper enriches the literature on psychological determinants in economic models for immigrants by studying the effect of the locus of control on immigrants' labor market outcomes.

3.3 Data and Variable Definitions

We use data from the 1999 and the 2007 wave of the German Socioeconomic Panel. The reason we use data of two different points in time is explained below in section 3.4. where the econometric strategy is explained. Despite some inaccuracies the GSOEP is a valuable dataset, especially rich through questions going beyond purely observable characteristics. It is particularly of interest for this study since it includes personality questions as well as migrant-related and detailed data on labor market returns and educational history of individuals. The sample consists of immigrants and natives aged 17-32 in 1999 (so 25-40

in 2007), not in education in 2007, with provided information on the dependent variables. Employment D and its determinants are measured in 2007. Psychometric measures and its determinants are measured in 1999. The sample size is 1812. There are 111 immigrants (6.1% of the sample) and 243 children of immigrants (13.4% of the sample). The German statistical office reports a percentage of 8.8% of "foreign population" (inhabitants of Germany with foreign nationality) in Germany in 2008 [9].

An immigrant is defined as *"foreign born with no German nationality at birth",* immigrants' children ("second generation") are defined as *"born in Germany with no German nationality at birth".*

To measure education levels I constructed three categories according to the ISCED[10] classification. ISCED 0-2 includes education up to the level of general elementary schooling and indicates a low education level, ISCED 3-4 includes "middle vocational schooling" and "vocational plus Abitur" and indicates a medium education level and ISCED 5-6 includes "higher vocational schooling" and "higher education" and indicates a high education level. Each category is controlled for by a dummy variable.

To take into account the different nationalities present in the sample I constructed three geopolitical nationality groups "EU15", "Central Europe and former Soviet Union" and "Turkey". Turkish immigrants are a large group among German non-nationals. A foreign language indicator takes the value one, if the only language spoken at home is the foreign language.

3.3.1 Measuring Personality: The Locus of Control

Personality measures have been developed by personality psychologists using self reported questionnaires with so-called psychometric questions. They found five factors - openness to experience, conscientiousness, extraversion, agreeableness, neuroticism. Motivation can be allocated to conscientiousness, which includes the facet "striving to achievement"[11]. Personality psychologists have found that personality is partly inherited and partly determined by the environment.

The psychometric questions to measure the locus of control are based on measures formulated by Rotter (1966). They are chosen from the following set of questions, present in the 1999 sample of the German Socioeconomic Panel:

1. How my life goes depends on me

[9] see http://www.statistik-portal.de/Statistik-Portal/de_jb01_jahrtab2.asp

[10] See UNESCO (2006): ISCED 1997 - International Standard Classification of Education, *www.uis.unesco.org.*

[11] See Borghans, Duckworth, Heckman and terWeel (2008).

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

2. Compared to other people, I have not achieved what I deserve

3. What a person achieves in life is above all a question of fate or luck

4. If a person is socially or politically active, he/she can have an effect on social conditions

5. I frequently have the experience that other people have a controlling influence over my life

6. One has to work hard in order to succeed

7. If I run up against difficulties in life, I often doubt my own abilities

8. The opportunities that I have in life are determined by the social conditions

9. Inborn abilities are more important than any efforts one can make

10. I have little control over the things that happen in my life

The answers can be "totally disagree", "slightly disagree", "slightly agree", "totally agree". I merge the first two categories to improve identification of the model, since the first two categories are characterized by low frequencies. Agreement with questions 1,4,6,9 is seen as an internal locus of control whereas agreement with questions 2,3,5,7,8,10 is seen as an external locus of control. The questions are chosen using the correlation matrix of the ten items. Five items display bivariate correlations with each other above 0.3, which is considered sufficiently large in this context. These five items are

2. Compared to other people, I have not achieved what I deserve

3. What a person achieves in life is above all a question of fate or luck

5. I frequently have the experience that other people have a controlling influence over my life

7. If I run up against difficulties in life, I often doubt my own abilities

10. I have little control over the things that happen in my life

3.3.1.1 The Relation between Economic Preferences and Personality Measures

Previous literature in economics has attempted to link psychometric questions to economic parameters. As mentioned above, Borghans, Heckman, Duckwoth and terWeel (2008) survey this emerging literature. They come to the conclusion that personality traits introduced both on an empirical and a theoretical level into economics can be fruitful for economic theory. According to the authors classical economic theory should incorporate the fact that economic preferences might be consequences of constraints imposed by cognitive skills and personality traits. They name the example that an agent with a high rate of time preference might be due to the fact that the agent cannot imagine the future. This ability to imagine

would be interpreted as a personal trait[12]. The authors see latent factor theory as a crucial connecting tool between psychology and economics. But the authors see as a major problem with this approach, especially in economics, the problem of reverse causality between personality traits and outcomes. A more self-confident person is prone to have a higher income, but a higher income also increases self-confidence. Borghans, Heckman et al (2008) name the work of Carneiro, Hansen and Heckman (2003) as a successful example for incorporating psychometric questions in an economic outcome model in a way that addresses the problem of endogeneity.

Borghans, Golsteyn Heckman and Meijers (2009) also contribute to this literature and study the psychological determinants of the risk aversion parameter, often found in economic theory. They study in particular the gender difference in risk aversion and find that people who are less agreeable, less neurotic and who have more ambition are less risk averse. They also find evidence for the hypothesis that the differences in risk aversion across gender are due to the fact that women differ in terms of their non-cognitive skills from men.

The theoretical economic literature also gives some economic insight applicable to the psychometric questions in this paper. For the questions 2 and 7 a working paper by Eeckhout and Weng (2009) provides some economic contents. In their labor learning model, neither workers nor firms know the worker's type, but they learn it by the wages a worker receives over time. Workers and firms can observe all wages received of all workers across time. This outcome process depends on a random error term. Such a model can explain why an individual would assume his abilities are of a low type if he receives only low wages. In question 7, "running up against difficulties in life", can be understood as receiving only lower wages. The fact that every worker can see the outcome of all other workers can explain why workers might think that others have achieved more than they, given their beliefs about their own abilities.

Another attempt to economically model the content of a set of psychometric questions - in this case the locus of control - was made by Bowles, Gintis and Osborne (2001a). They develop a theoretical model to explain the advantage for an employer to employ a motivated worker by setting up a signalling model. They interpret the locus of control as an employee's preference, which reduces the employer's cost to induce the employee's effort. This simply means that a worker, who believes more in his own success is more motivated to induce effort. This could also be interpreted that a worker with intrinsic motivation will not need an external source to make him provide an effort. This is a desirable trait of a worker for

[12] As adressed in our econometric strategy below, this trait could also be due to the family background and previous experiences.

an employer and the authors call such a trait an "incentive enhancing preference". Bowles, Gintis and Osborne (2001a) use a principal-agent model to interpret the relationship between incentive-enhancing preferences and earnings in an economic way. In the model the employer (the principal) and the employee (the agent) cannot contract on effort. The employer has an imperfect measure of effort. The employee has an effort-dependent belief on the termination of the contract. An incentive-enhancing preference is in this model an employee's attribute, which makes him work harder at every wage level. As an example of an incentive-enhancing preference in this model the authors name the locus of control. An individual with an external locus of control is a fatalist person who does not believe that effort can change any outcomes in his life. In Bowles' et al (2001a) model fatalism would enter the model via lowering the worker's belief on the effectiveness of effort on lowering the probability of termination of the contract. In the model fatalism lowers the marginal subjective benefit to exerting effort in the agent's first order condition. By entering the first order condition greater fatalism lowers the agent's optimal choice of effort.

The reviewed literature in the paragraph shows that there is a growing recognition in economics and translation into economic concepts of psychological concepts and psychometric measures.

3.4 Econometric Strategy

As briefly noted in the data description in section 3.3., we measure the labour market outcomes and its determinants (apart from the latent factor) in 2007, whereas the latent factor and its determinants are measured in 1999. The reason for this is to address the reverse causality that could exist between the latent factor and the labour market outcome. The idea is to measure the latent factor - the Rotter index - at the earliest time possible given the data, which is the time of finalizing the education and early labour market entry. So we choose to take those that are 17-32 years old in 1999 into the sample. Labour market outcomes are then measured eight years later in 2008. We adopt this strategy to address the issue of reverse causality given the data available. The main GSOEP dataset is only available for individuals older than 16[13].

As a first step we estimate the model in a traditional way: we treat the Rotter index as exogenous and estimate a psychometric model separately from an economic outcome model. As a second step we endogenize the Rotter index, which allows us to examine the effects of the factors determining the locus of control. This approach allows a comparison between the traditional method and an integrated methodology which allows to address the problem

[13] GSOEP subsamples exist for youth but include only a restricted set of variables.

of endogeneity and to assess the effect of determinants on the locus of control. In the next sections I will first outline the two methodologies - the traditional one and the integrated one and then proceed to the interpretation of the results.

3.4.1 Exogenous Personality : Two step estimation procedure

3.4.1.1 The Personality Model

As a first step, we treat the Rotter index as exogenous and we estimate two models separately. The first model is a classic factor model. Factor models have been developed in psychology to measure intelligence[14]. Later factor models were also used to measure other personality traits, in political science for measuring concepts and in financial economics to measure latent concepts which influence financial markets.

The main idea of factor models is to use a set of measures for the concept "intelligence", "discipline","peace" or "beliefs on the stock market" and to divide the joint variation among these measures into a common part θ and a random part ε and to estimate the common part θ and its effect on the measures, indicated by α. θ indicates in this paper the locus of control, which is measured using a set of questions related to the locus of control[15]. The model is a simultaneous equation model of the five psychometric questions above. Each psychometric question is modelled as an ordered probit model. All five questions are assumed to depend on a latent factor θ, the locus of control, and an independent random error term ε^M. The psychometric questions all depend differently on the latent factor - each question has a different factor loading α^M,which can be interpreted as a coefficient of the latent factor in the regression of M on θ.The model is estimated with a maximum likelihood methodology, implemented in STATA. The model takes the form:

$$\begin{aligned} M_{99} &= \{1,2,3\} \\ M^*_{99} &= \alpha^M \theta_{99} + \varepsilon^M_{99} \end{aligned}$$

where the subscripts denote the year of measurement of the variable.

We include five items in our model. Writing the equations for each items (and dropping

[14] See Spearman (1904).
[15] These questions have been developed by Rotter (1966).

the subscript for the year the variable is measured in) gives

$$\begin{aligned} M_1 &= \{1,2,3\} \\ M_1^* &= \alpha^{M_1}\theta + \varepsilon^{M_1} \\ M_2 &= \{1,2,3\} \\ M_2^* &= \alpha^{M_2}\theta + \varepsilon^{M_2} \\ M_3 &= \{1,2,3\} \\ M_3^* &= \alpha^{M_3}\theta + \varepsilon^{M_3} \\ M_4 &= \{1,2,3\} \\ M_4^* &= \alpha^{M_4}\theta + \varepsilon^{M_4} \\ M_5 &= \{1,2,3\} \\ M_5^* &= \alpha^{M_5}\theta + \varepsilon^{M_5} \end{aligned}$$

Parametric Identification of Factor Models Here I give a brief outline of the identification of factor models. Factor models take the form of the measurement equation above:

$$M^* = \alpha^M\theta + \varepsilon^M$$

Consider M^* to be computable.

The identification of factor models is based on the covariance matrix of the items:

$$cov(M^*) = \Lambda\Sigma_f\Lambda' + \Omega_e$$

where

$$\begin{aligned} \theta &\perp \varepsilon^M \\ \varepsilon^M &\sim N(0,1) \end{aligned}$$

Λ - matrix of factor loadings α^M

Σ_f - variance-covariance matrix of the factors

Ω_e - diagonal matrix of "uniqueness"-variances of ε^M

K - number of factors θ

L - number of items M

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

The goal is to identify KxL factor loadings Λ and K variances of factors Σ_f . The elements of $cov(M)$ are observable and the elements of Ω_e are determined by our distributional assumption on ε^M. So we can identify the unobservable elements Λ and Σ with the $(L(L-1)/2)$ observable off-diagonal elements of $cov(M)$.So,we need that

$$L(L-1)/2 \geq (L \times K) + K.$$

The number of unique terms in $cov(M)$ needs to be equal to or larger than the number of factor variances and factors. In our case $K = 1$ and $L = 5$. So we have

$$\begin{aligned} 5 * 4/2 &\geq 5 + 1 \\ 10 &\geq 6 \end{aligned}$$

3.4.1.2 The Employment Model

The second model is an employment model. The latent factor θ, estimated through the model above, is treated as an additional explanatory variable in the employment equation. The model takes the form:

$$\begin{aligned} D_{07} &= \{0, 1\} \\ D^*_{07} &= \beta^D_0 + \alpha^D \theta_{99} + \beta^D X_{07} + \varepsilon^D_{07} \end{aligned}$$

3.4.2 Endogenizing Personality: Simultaneous Equation Model

The models above treats the personality measure, the locus of control, as exogenous. As a next step, we endogenize the locus of control - firstly to address its endogeneity in an employment equation and secondly to find out, how the locus of control is determined. Especially, we are interested in whether immigrants and their children have different positions on the locus of control scale. Additionally, the methodology, we use for this, allows to estimate all parameters and the locus of control at the same time. This avoids treating the measure of the locus of control as an observed quantity as we did above.

The model is a linear parametric simultaneous equation model with an embedded factor model structure, as described above. The simultaneous equation model contains the equations for the economic outcome D and for the measures M. In this paper the latent concept "locus of control" is endogenized and so I add another equation in the simultaneous equation model to determine θ

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

The model then takes the following form:

$$\begin{aligned} D_{07} &= \{0,1\} \\ D^*_{07} &= \beta^D_0 + \alpha^D \theta_{99} + \beta^D X_{07} + \varepsilon^D_{07} \\ M_{99} &= \{1,2,3\} \\ M^*_{99} &= \alpha^M \theta_{99} + \varepsilon^M_{07} \\ \theta_{99} &= \gamma W_{99} + \varepsilon^\theta_{99} \end{aligned}$$

where D is an employment indicator, M signifies psychometric measures for the locus of control. Since M and D are categorical variables we need to impose a probit structure on the variables, so D^* and M^* indicate the latent underlying variables for the probit models for M and D. X comprises the observable variables (called *direct effects*). W comprises the observable variables (called *indirect effects*) "age", "gender", "immigrant", "religion important", "father upper secondary education", "mother upper secondary education", "father education missing", "mother education missing", "employment in 99", "still in education in 99".

3.4.2.1 Identification Assumptions

The identification strategy is parametric and is described in detail in section 2.2.1.1. At this point we mention the assumptions - adapted to the particular formulation of the model in this chapter - and refer to the discussion in section 2.2.1.1.

First we need to make assumptions concerning the factor analytical part of the model:

$$V(\theta) = 1$$

We remind the reader that - as shown in section 2.2.1.1. - the distribution and the mean of θ follow from the model and the distributional assumption below on the error term ε_θ.

$$\theta \sim N(\gamma W, 1)$$

Next we impose distributional restrictions on the error terms

$$\begin{aligned} \varepsilon_\theta &\sim N(0,1) \\ \varepsilon_D &\sim N(0,1) \\ \varepsilon_M &\sim N(0,1) \end{aligned}$$

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

In addition, we impose the following (conditional) independence conditions:

$$\begin{aligned} W &\perp \varepsilon_\theta \\ \theta &\perp \varepsilon_M|W \\ \theta &\perp \varepsilon_D|W \\ X &\perp \theta|W \\ \theta &\perp \varepsilon_D \\ D &\perp M|\theta \\ M_i &\perp M_j|\theta \quad \forall i \neq j \end{aligned}$$

Finally, we impose the normalization that, for all tricategorical items the cutpoint between the first and the second category is

$$c_{1j} = 0 \quad \forall j$$

3.4.2.2 Estimation: The Gibbs Sampler

The likelihood function of the model under the assumption of independently and identically distributed observations is given by

$$\begin{aligned} & \prod_{i=1}^{N} f(M_i, D_i, M_i^*, D_i^*, \theta i|X_i, W_i, \alpha, \beta, \gamma, c) \\ = & \prod_{i=1}^{N} f(M_i^*, D_i^*, \theta_i|X_i, W_i, \alpha, \beta, \gamma, c) \prod_{i=1}^{N} f(M_i, D_i|\theta_i, M_i^*, D_i^*, X_i, W_i, \alpha, \beta, \gamma, c) \\ = & \prod_{i=1}^{N} f(M_i^*, D_i^*, \theta_i|X_i, W_i, \alpha, \beta, \gamma, c) \prod_{i=1}^{N} f(M_i, D_i|c) \end{aligned}$$

where the factor loadings are written as $\alpha = (\alpha^M, \alpha^D)$ and the coefficients as $\beta = \beta^D$.The first simplification follows from exploitation of the product rule. The second step follows from the fact that ordinal responses are solely determined by the underlying variables D_i^* and M_i^* and by the cutpoints c. We can factor out the likelihood function $f(M_i^*, D_i^*, \theta_i|X_i, W_i, \alpha, \beta, \gamma, c)$ into $f(M_i^*, \theta_i|.)f(D_i^*, \theta_i|.)$ due to the conditional independence assumptions above. The likelihood functions of D_i^* and M_i^* written separately are

$$\prod_{i=1}^{N}[f(M_i^*,\theta_i|\alpha,\gamma,c,M_i,W_i)\{\sum_{k_M=1}^{K_M}1(M_i = k_M)1(c_{k_M-1}<M_i^*<c_{k_M})\}]$$
$$\prod_{i=1}^{N}[f(D_i^*,\theta_i|\alpha,\beta,\gamma,D_i,X_i,W_i)\{\sum_{k_D=1}^{K_D}1(D_i = k_D)1(c_{k_D-1}<D_i^*<c_{k_D})\}]$$

Each of the factors $f(M_i^*,\theta_i|.)$ and $f(D_i^*,\theta_i|.)$ needs to be multiplied by two indicators - an indicator which equals one if the observation M_i (D_i) falls in category k_M (k_D) and an operator indicating that M_i^* (D_i^*) must fall between the two cutpoints c_{k_M-1} (c_{k_D-1}) and $c_{k_M}(c_{k_D})$ according to its category .

θ is unobservable and will be estimated. To make the mechanism by which θ_i influences M_i^* and of D_i^* perspicuous we integrate out θ_i and obtain the distributions of M_i^* and D_i^* conditional on the parameters of the model and on the data.

$$f(M_i^*|\alpha,c,\gamma,M_i,W_i) = \int_\theta f(M_i^*|\alpha,c,\theta_i,M_i)f(\theta_i|\gamma,W_i)d(\theta_i)$$
$$f(D_i^*|\alpha,\beta,\gamma,D_i,X_i,W_i) = \int_\theta f(D_i^*|\alpha,\beta,c,\theta_i,D_i,X_i)f(\theta_i|\gamma,W_i)d(\theta_i)$$

It becomes obvious that the likelihood function of the model is a high-dimensional integral, which cannot be solved analytically and needs to be solved by numerical methods. Markov Chain Monte Carlo[16] methods provide a way to estimate the parameters of interest by sampling from the integral. The main advantage of the Gibbs sampler is its relative computational ease.

The Gibbs sampler is a Bayesian method. The Bayesian paradigm specifies statistical models as a posterior joint distribution, composed of the two elements prior distribution and likelihood function. The prior distribution contains the beliefs of the researcher about the parameters before taking into account the information in the data. The prior is combined with the likelihood function, which contains the information of the data. The posterior joint distribution is obtained by simply multiplying the priors with the likelihood and it can be written as

[16] The Gibbs sampler and Bayesian statistics are assessed in chapter one.

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

$$\begin{aligned} & f(\beta, \alpha, \gamma, \theta_i, M^*, D^*, c|M, D, X, W) \\ \propto \quad & f(\beta)f(\alpha)f(\gamma)f(c)\prod_{i=1}^{N} f(M_i, D_i, M_i^*, D_i^*, \theta_i|X_i, W_i, \alpha, \beta, \gamma, c) \end{aligned}$$

where $f(\beta)f(\alpha)f(\gamma)f(c)$ are the priors for the coefficients of X, the factor loadings, the coefficients of W and the cutpoints.

The Gibbs sampler is an algorithm which samples from this joint posterior distribution in a sequential way. The idea of the Gibbs sampler is to sample one of the elements among $M_i^*, D_i^*, \beta, \alpha, \gamma, c$ and θ at a time, conditioning on the last sampled values for the remaining elements and on the data. This procedure is equivalent to sampling from a set of conditional distributions sequentially. Each conditional distribution is a conditional posterior distribution of a parameter value given the last sampled values of the other parameters and the data. These conditionals - each of them constitutes one step of the Gibbs sampling algorithm - are called "full conditionals". The closed form of the full conditionals follows from the properties of the model. After a sufficient amount of iterations, the algorithm converges under a set of regularity conditions[17] and the sampled values are samples from the true posterior[18]. The algorithm for the model in this paper ran for 100 000 iterations and convergence statistics do not indicate that the algorithm has not converged. In the following I derive the full conditionals of the model.

First a value is sampled from the posterior conditional distribution (or full conditional) of the latent underlying variables, then from the posterior conditional distribution of the factor loadings and so forth. For the second iteration the same procedure is repeated, conditioning on the sampled values from the first iteration. The very first iteration starts with a set of specified initial values. The algorithm is not sensitive to the choice of the starting values.

3.4.2.3 The Posterior Conditional Distribution of the Latent Underlying Variables

Albert and Chib (1993) propose a data augmentation procedure to sample latent underlying variables in a threshold model. It follows from their work that the full conditional for the latent underlying variable of the binary response is

[17] These conditions are assessed in chapter one.
[18] For the theory MCMC algorithms and on the Gibbs sampler, see Robert and Casella (2004).

$$f(D^*|\alpha^D, \beta^D, \theta, D, X) \propto$$
$$\prod_{i=1}^{N} f(D_i^*|\beta^D X_i^D + \alpha^D \theta_i, 1)\{\sum_{k_D=1}^{K_D} 1(D_i = k_D) 1(c_{k_D-1} < D_i^* < c_{k_D})\}$$

where $\alpha^D, \beta^D, \theta$ signify the last sampled values from the previous iteration of the algorithm. It follows from the normality assumptions on θ and ε that $f(D_i^*|\theta_i, \alpha, \beta, D_i, X_i)$ is normally distributed - with mean $\beta^D X_i^D + \alpha^D \theta_i$ and $V(D_i^*)$ normalized to one as indicated above.

The latent underlying variable is distributed as the following truncated normal distributions:

$$D_i^*|\alpha, \beta, \theta, D, X \tilde{} TN_{(-\infty,0)}(\beta^D X_i^D + \alpha^D \theta_i, 1) \text{ if } D_i = 0$$
$$D_i^*|\alpha, \beta, \theta, D, X \tilde{} TN_{(0,\infty)}(\beta^D X_i^D + \alpha^D \theta_i, 1) \text{ if } D_i = 1$$

Similarly, the full conditionals for the polytomous variables are

$$f(M^*|\alpha, \theta, c, M, X) \propto$$
$$\prod_{i=1}^{N} f(M_i^*|\alpha^M \theta_i, 1)\{\sum_{k_M=1}^{K_M} 1(M_i = k_M) 1(c_{k_M-1} < M_i^* < c_{k_M})\}$$

The latent underlying variables of the polytomous items is distributed as the following truncated normal distribution:

$$M_i^*|\alpha, \theta, c, M, X \sim TN_{(c_{k_M-1}, c_{k_M})}(\alpha^M \theta_i, 1)$$

3.4.2.4 The Posterior Conditional Distribution of the Factor Loadings

The full conditional for the factor loadings for D and M can be written as[19]

$$f(\alpha^D|\beta, \theta, D, X, D^*) \propto f(\alpha^D) \prod_{i=1}^{N} f(D_i^*|\beta^D X_i^D + \alpha^D \theta_i, 1)$$
$$f(\alpha^M|\theta, M, X, M^*) \propto f(\alpha^M) \prod_{i=1}^{N} f(M_i^*|\alpha^M \theta_i, 1)$$

[19] As above β and θ denote the last sampled values.

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

where we choose normal priors $f(\alpha^D) = N(0,1)$ and $f(\alpha^M) = N(0,1)$.If we rewrite the equation for D^* and M^*as

$$\begin{aligned} D_i^* - \beta^D X_i^D &= \alpha^D \theta_i + \varepsilon_i^D \\ M_i^* &= \alpha^M \theta_i + \varepsilon_i^M \end{aligned}$$

we can treat it as a normal regression model and derive for M and D

$$\begin{aligned} \alpha^M | \theta_i, M_i, M_i^* &\sim N\left[(\theta_i'\theta_i + 1)^{-1}\theta_i'(M_i^*), (\theta_i'\theta_i + 1)^{-1}\right] \\ \alpha^D | \beta, \theta_i, D_i, X_i, D_i^* &\sim N\left[(\theta_i'\theta_i + 1)^{-1}\theta_i'(D_i^* - \beta^D X_i^D), (\theta_i'\theta_i + 1)^{-1}\right] \end{aligned}$$

3.4.2.5 The Posterior Conditional Distribution of the Direct Coefficients

Similarly to the procedure for the factor loadings, we can write the model as

$$D_i^* - \alpha^D \theta_i = \beta^D X_i^D + \varepsilon_i^D$$

For the coefficients, we choose to set diffuse priors as well. The full conditionals for the intercepts are, according to Albert and Chib (1993,p.671)

$$\beta^D | \alpha, \theta_i, D_i, X_i, D_i^* \sim N\left[(X_i'X_i)^{-1}X_i'(D_i^* - \alpha^D \theta_i^D), (X_i'X_i)^{-1}\right]$$

3.4.2.6 The Posterior Conditional Distribution of the Cutpoints

We assume a uniform prior for the cutpoints and can write for the full conditionals for the polytomous responses

$$c^M | \alpha, \theta, M, M^* \sim unif \begin{bmatrix} \max\{\max\{M_i^* : M_i = k_M\}, c_{M-1}\}, \\ \min\{\min\{M_i^* : M_i = k_{M+1}\}, c_{M+1}\} \end{bmatrix}$$

3.4.2.7 The Posterior Conditional Distribution of the Latent Factors

Similarly as for the procedure for coefficients and factor loadings, we can rewrite the model as

$$\begin{aligned} D_i^* - \beta^D X_i^D &= \alpha^D \theta_i + \varepsilon_i^D \\ M_i^* &= \alpha^M \theta_i + \varepsilon_i^M \end{aligned}$$

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

and treat it as a normal regression model,where θ_i is the parameter to be estimated. Carneiro, Hansen and Heckman (2003) specify a mixture of normals as the prior for the latent factors. We treat the latent factors as endogenous depending on γW_i. We treat θ_i in the same way as M_i^* and D_i^* for which the priors are implicitly determined by the prior distributions of the other parameters and by the assumptions on the distribution of ε_i^M and ε_i^D. The prior of θ_i is therefore implicitly determined by the priors of the other parameters of the model and by the assumptions on the distributions of ε_i^M,ε_i^D and ε_i^θ.

We can derive the full conditional for the latent factor as:

$$\begin{aligned} & f(\theta|\beta,\alpha,c,\gamma,X,W,D^*,M^*) \\ \propto & \prod_{i=1}^{N} f(M_i^*|\alpha^M\theta_i,1)f(D_i^*|\beta^D X_i^D+\alpha^D\theta_i,1) \end{aligned}$$

We do not need to condition on D and M since they are implicitly known through D^* and M^* and c.Our dependent variables are ordinal and for identification reasons their variances and error variances have been set to one.

The posterior conditional distribution of θ_i is given by:

$$\begin{aligned} & \theta_i|\beta,\alpha,\gamma,c,X_i,W_i,Di^*,M_i^* \\ \sim & N\left[\begin{array}{c} \gamma W_i+(\alpha^{D\prime}\alpha^D+\alpha^{M\prime}\alpha^M+1)^{-1} \\ (\alpha^{M\prime}(M_i^*-\alpha^{M\prime}\gamma W_i)+\alpha^D(D_i^*-\beta^D X_i^D-\alpha^D\gamma W_i)), \\ I-\alpha^{D\prime}(\alpha^{D\prime}\alpha^D+\alpha^{M\prime}\alpha^M+1)^{-1}\alpha^D \\ -\alpha^{M'}(\alpha^{D\prime}\alpha^D+\alpha^{M\prime}\alpha^M+1)^{-1}\alpha^M \end{array}\right] \end{aligned}$$

3.4.2.8 The Posterior Conditional Distribution of the Indirect Coefficients

The posterior we sample from can be written as

$$\begin{aligned} & f(\gamma|\theta,W) \\ \propto & f(\gamma)f(\theta|\gamma,W) \end{aligned}$$

The model for the latent variable is

$$\theta=\gamma W+\varepsilon^\theta$$

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

We assume a diffuse prior for the coefficient γ. Similar to the procedures above we get:

$$f(\gamma|\theta, W) \sim N((W'W)^{-1}W'\theta), (W'W)^{-1})$$

3.5 Results

Table 3.1 shows the results of a simple equation for employment including two dummy variables for immigrants and for the second generation. The base category are native-born who have the German nationality at birth. All coefficients display the expected signs. Being an immigrant lowers the employment probability significantly. This effect is attenuated for the second generation whereas the coefficient for the second generation is not significant. Being married does not significantly increase the probability of being employed whereas having children does[20]. The table shows results from two different estimation methods - firstly, from the Markov Chain Monte Carlo methodology, which is used later for the incorporation of the latent factor and secondly, the traditional Maximum Likelihood estimator for probit models. The results show that both methods give quite similar results. This fact is taken as evidence for the correct implementation of the Markov Chain Monte Carlo method since the assumptions used for both methods are the same. It can also be interpreted as evidence for the fact that the priors chosen are not very informative and that the MCMC estimator takes the information given in the data more strongly into account - in other words the data is informative.

In table 3.2 the results for a model taking into account the ethnic background and the language spoken at home. I split the sample into four ethnic groups - four dummy variables control for Turkish, central European (this group includes the former Soviet Union), EU15 (including Switzerland and the US) and German nationalities. The base category are all German-born with German nationality at birth[21]. The results for the control variables are similar to the table 3.1. The ethnic variables show that only the Turkish first and second generation have a significant disadvantage compared to native Germans. It must be noted however, that the sample sizes for the different ethnic groups are small and this might affect the significance of the variables.

[20] It would be of interest to split the sample into men and women since it could certainly be the case that the coefficient for marital status differs largely between men and women and therefore renders the coefficient insignificant once taking the whole sample. But here the interest lies in the immigrant population and I consider the immigrant and the second generation samples as too small to be able to split the sample.

[21] There is no group for immigrants with German nationality since there are none in the sample.

β^D	MCMC	ML
Intercept	-1.00	-0.481
	(-1.99,-0.04)	(0.340)
Age	0.03	0.036
	(0.01,0.05)	(0.009)
Gender	-0.56	-0.573
	(-0.70,-0.421)	(0.072)
Immigrant	-0.40	-0.406
	(-0.67,-0.13)	(0.139)
Second generation	-0.16	-0.115
	(-0.34,0.036)	(0.098)
Low education	-0.40	-0.411
	(-0.58,-0.21)	(0.096)
High education	0.31	0.293
	(0.14,0.47)	(0.085)
Marital status	0.08	0.078
	(-0.07,0.24)	(0.081)
Children under 16	0.23	0.260
	(0.08,0.37)	(0.077)

Table 3.1: Estimates of the Employment Equation: $D^* = \beta X + \varepsilon_D$, estimated by MCMC and ML

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

β^D	ML	MCMC
Intercept	-0.57	-1.34
	(0.342)	(-2.31,-0.31)
Age	0.04	0.04
	(0.000)	(0.02,0.06)
Gender	-0.58	-0.59
	(0.072)	(-0.72,-0.44)
Turkish immigrant	-0.65	-0.65
	(0.210)	(-1.06,-0.23)
Central European immigrant	-0.07	-0.06
	(0.279)	(-0.60,0.50)
EU15 immigrant	0.47	0.50
	(0.352)	(-0.18,1.23)
Turkish second generation	-0.27	-0.45
	(0.177)	(-0.87,-0.02)
Central European second generation	0.31	0.32
	(0.256)	(-0.28,0.95)
EU15 second generation	0.18	0.16
	(0.234)	(-0.30,0.64)
German second generation	-0.08	-0.13
	(0.135)	(-0.38,0.12)
Immigrant foreign language spoken at home	-0.52	-0.53
	(0.332)	(-1.19,0.12)
Second generation foreign language spoken at home	-0.26	-0.25
	(0.330)	(-0.89,0.41)
Low education	-0.36	-0.37
	(0.097)	(-0.56,-0.17)
High education	0.28	0.28
	(0.081)	(0.11,0.45)
Marital Status	0.10	0.10
	(0.08)	(-0.06,0.25)
Children under 16	0.26	0.26
	(0.773)	(0.10,0.41)

Table 3.2: Estimates of the Employment Equation: $D^* = \beta X + \varepsilon_D$, estimated by MCMC and ML

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

3.5.1 Adding Personality

Table 3.3 to 3.5 show the results for a simple model of employment - for natives, immigrants and their children, adding a measure of the locus of control. The employment equation in table 3.3. includes the controls age, gender, educational indicators, marital status, number of children and indicators for the migration generation as well as the locus of control. The locus of control equation in table 3.5. includes those variables included also in the employment equation, which could be correlated with the locus of control. It is necessary for identification - as shown in chapter two - that the locus of control θ and the determinants X of employment are not correlated. We assume in line with Matzkin (2003,2007) that once we control for those elements which could cause a correlation between θ and X,they are uncorrelated. These control variables, which determine directly employment and which are likely to be also correlated with the locus of control are age, gender, the immigrant status and generation and education. Marital status and the number of children in 2007 are assumed not to determine the latent factor measured in 1999. As shown in table 3.5. the locus of control equation includes in addition also parental education, an indicator whether religion is important and the time since immigration to Germany measured in 1999. These variables are assumed to have a direct effect on the locus of control but only an indirect (via the locus of control) effect on the employment probability. Whether religion is important in 1999 we see as a determinant for the locus of control (Kahoe 1974) but not as a determinant of employment. The variables parental education and time already stayed in Germany in 1999 affect the locus of control in 1999 but do not affect employment in 2007 directly. The reasoning behind this way of modelling is that the time stayed in Germany and parental education can only have a positive effect on employment when they enhance motivation. Gender can have a direct effect on employment due to possible gender discrimination. Age (and thereby experience) due to the fact that employers look at the age and experience of an individual directly when deciding on employment. Immigrant status can have a direct effect due to possible discrimination. Marital status and the number of children have an impact on whether an individual needs to work - the classical reasoning is that a married man with several young children is more likely to need a job to feed his family. Parental background however has an effect on the shaping of personality and therefore on the shaping of the locus of control. The same is true for the time stayed in Germany - it can contribute to integration and shapes an individual's character.

The results in table 3.3 are quite similar to those of table 3.1. The coefficient for immigrants is slightly less negative than without controlling for the locus of control and the coefficient for the second generation is much less negative but still insignificant. Again, being

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

an immigrant is a disadvantage on the labour market and the disadvantage is attenuated for the second generation. The last row in table 3.3 show the results for the locus of control. They show that having a more internal locus of control has a positive and significant effect on the employment probability. We can see that a 2.5 σ locus of control units can compensate for being an immigrant and 1 for being a second generation immigrants. About four σ locus of control units can compensate for having a low educational attainment level as opposed to a medium one.

The three columns in table 3.3 refer to three different ways of estimating the model - one Maximum Likelihood method and two MCMC methods. The Maximum Likelihood method is a two-step method whereas the MCMC methodology allows to estimate all parameters simultaneously. The former way of estimating a model with latent variables is easy to implement since most software includes a routine to estimate a conventional factor model and a routine to estimate a probit model. Both models - the probit model for the outcome equation and the factor model - can be estimated using maximum likelihood procedures[22]. A two-step methodology has the disadvantage, that the latent factor is treated as an observable variable in the second stage. This is a less efficient method than a method estimating all parameters simultaneously and taking into account that the latent variable is an estimated and not an observable entity. On the other hand, if the latent factor is estimated in a wrong way, any mistake is carried on to the estimation of the remaining parameters. My results show that both estimation methods render highly similar results - even for the coefficient of the latent factor.

As outlined above, the locus of control is treated here as a variable that is determined by socioeconomic conditions and possibly partly also by genetical heritage. The second and third column show results for a model in which the locus of control is estimated without assuming that it is determined by other variables and for a model in which the locus of control is estimated assuming that it is determined by socioeconomic variables, respectively. The table shows that the results do not differ greatly.

Table 3.4 shows the estimations of the factor loadings. They are positive for all items and using all three ways of estimating the model. They differ in size depending on the estimation methodology used and on whether the latent factor is treated as exogenous or as endogenous. In table 3.5 I show the coefficients of the determinants of the locus of control. The results show that immigrants have a much lower level of beliefs that they can influence

[22] A maximum likelihood method for estimating a factor model is Rao's canonical-factor method, which is based on maximizing the determinant of the correlation matrix of the items by seeking the highest canonical correlation with the items. This methodology is based on the assumption that the factors are continuous. Certainly this is a less accurate methodology but for the mere purpose of comparison to the methodology used throughout my dissertation I assume the level of accuracy sufficient.

the outcomes of their lives; they seem to be more fatalistic. The same is true for the second generation but the effect is attenuated. Age and the duration of stay in Germany have a small, significantly positive effect on the locus of control. Mother's education - interestingly, as opposed to father's education - has a positive and significant effect on the locus of control. Education does not have a significant effect on the locus of control, but the fact of still being in education does.

The outlined results suggest that motivation matters for everyone and that immigrants can overcome their gap by motivation. But immigrants have a *double disadvantage* since they have less motivation than natives and more motivation would actually help them find a job.

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

β^D,α^D	ML	MCMC exog	MCMC endog
Intercept	-0.46	-1.38	-0.75
	(0.341)	(-2.34,-0.29)	(-1.71, 0.23)
Age	0.04	0.04	0.03
	(0.009)	(0.02,0.05)	(0.01,0.05)
Gender	-0.57	-0.58	-0.57
	(0.072)	(-0.71,-0.43)	(-0.70,-0.42)
Immigrant	-0.37	-0.37	-0.34
	(0.141)	(-0.65,-0.09)	(-0.62,-0.05)
Second generation	-0.09	-0.09	-0.11
	(0.098)	(-0.29,0.10)	(-0.31,0.08)
Low education	-0.39	-0.40	-0.40
	(0.095)	(-0.59,-0.20)	(-0.59,-0.21)
High education	0.27	0.27	0.27
	(0.085)	(0.10,0.45)	(0.10,0.44)
Marital status	0.07	0.07	0.02
	(0.081)	(-0.09,0.23)	(-0.13,0.18)
Children under 16	0.26	0.26	0.15
	(0.077)	(0.11,0.42)	(0.001,0.29)
Locus of control	0.12	0.17	0.13
	(0.041)	(0.04,0.24)	(0.05,0.21)

Table 3.3: Estimates of the Employment Equation: $D^* = \alpha\theta + \beta X + \varepsilon_D$, estimated by MCMC and ML

α^M	ML	MCMC exog	MCMC endog
Not achieved what I deserve	0.49	0.99	0.62
		(0.88,1.10)	(0.56,0.69)
Achievements are question of luck	0.44	0.72	0.44
		(0.64,0.80)	(0.39,0.49)
Other people influence my life	0.55	1.17	0.72
		(1.06,1.30)	(0.65,0.79)
Doubt my abilities	0.52	0.98	0.58
		(0.88,1.08)	(0.53,0.65)
Little control over my life	0.67	1.84	1.37
		(1.60,2.09)	(1.19,1.56)

Table 3.4: Estimates of the Psychometric Question Equations: $M^* = \alpha\theta + \varepsilon_M$

	γ	Bayesian confidence intervals
Age	0.04	0.03, 0.04
Gender	-0.03	-0.15, 0.08
Immigrant	-0.80	-1.38, -0.18
Second generation	-0.23	-0.44, -0.00
Religion important	-0.03	-0.17, 0.10
Low education in 1999	-0.03	-0.17, 0.10
High education in 1999	0.16	-0.02, 0.35
Father highly educated	0.00	-0.19,0.19
Mother highly educated	0.35	0.11,0.59
In education in 1999	0.22	0.05, 0.37
Time stayed in Germany	0.04	0.00, 0.07

Table 3.5: Estimates of Determinants of the Locus of Control: $\theta = \gamma W + \varepsilon_\theta$, estimated by MCMC

In table 3.6 to 3.8 I add nationalities and an indicator of whether German is spoken at home. The control variables do not change much in size or sign. The results show that only Turkish immigrants have a significant disadvantage on the German labour market once controlling for the locus of control. It should be noted that, as mentioned above, the sample sizes for the separate ethnic groups are small. A Turkish immigrant can compensate his disadvantage on the labour market by six σ units of belief in being able to determine his success. As in table 3.2, speaking only the foreign language at home does not cause a significant disadvantage for immigrants or for the second generation. The locus of control still has a positive and significant effect. Table 3.6 displays again three columns for the three ways of estimating the model - by Maximum Likelihood, by MCMC assuming a latent factor not determined by other variables and assuming a latent factor determined by socioeconomic controls. Again, the results are largely similar across the methods.

Table 3.7 show the results for the factor loadings which are all positive and significant. Table 3.8 shows the results for the determinants of the locus of control, when adding nationalities and an indicator of whether German is spoken at home. Age has a small and significantly positive effect on the locus of control and mother's education seems to be an important determinant. Turkish immigrants as well as the Turkish second generation have a significantly lower belief in being able to determine their own success.

Again, the results show that - especially the Turkish - immigrants and their children have a double disadvantage on the labour market: they are disadvantaged in terms of employment and additionally they lack in belief to be able to do something to be successful in life - a skill which matters on the labour market.

β^D,α^D	ML	MCMC exog	MCMC endog
Intercept	-0.55	-1.43	-1.41
	(0.343)	(-2.41,-0.37)	(-2.41,-0.39)
Age	0.04	0.04	0.04
	(0.009)	(0.02,0.06)	(0.02,0.06)
Gender	-0.58	-0.59	-0.58
	(0.072)	(-0.72, -0.44)	(-0.72,-0.44)
Low education	-0.36	-0.36	-0.36
	(0.098)	(-0.55,-0.16)	(-0.55,-0.17)
High education	0.26	0.26	0.26
	(0.086)	(0.09,0.43)	(0.09,0.42)
Turkish immigrant	-0.61	-0.63	-0.60
	(0.211)	(-1.05,-0.21)	(-1.02,-0.18)
Central European Immigrant	-0.06	-0.05	-0.06
	(0.282)	(-0.61,0.52)	(-0.60,0.51)
EU15 immigrant	0.51	0.55	0.56
	(0.355)	(-0.17,1.26)	(-0.15,1.26)
Turkish second generation	-0.25	-0.44	-0.41
	(0.177)	(-0.87,0.00)	(-0.85,0.01)
Central European second generation	0.30	0.32	0.32
	(0.255)	(-0.30,0.94)	(-0.29,0.92)
EU15 second generation	0.19	0.19	0.18
	(0.235)	(-0.29,0.66)	(-0.28,0.66)
German second generation	-0.07	-0.11	-0.11
	(0.135)	(-0.37,0.14)	(-0.36,0.15)
Immigrant foreign language spoken at home	-0.47	-0.48	-0.47
	(0.334)	(-1.14, 0.19)	(-1.13,0.19)
Second generaion foreign language spoken at home	-0.18	-0.14	-0.16
	(0.333)	(-0.81,0.51)	(-0.82,0.49)
Marital status	0.09	0.09	0.09
	(0.082)	(-0.07,0.25)	(-0.08,0.24)
Children under 16	0.26	0.26	0.26
	(0.077)	(0.11,0.41)	(0.11,0.41)
Locus of control	0.10	0.10	0.10
	(0.042)	(0.02,0.23)	(0.02,0.18)

Table 3.6: Estimates of the Employment Equation: $D^* = \alpha\theta + \beta X + \varepsilon_D$, estimated by MCMC and ML

α^M	ML	MCMC exog	MCMC endog
Not achieved what I deserve	0.49	0.99	0.62
		(0.89,1.09)	(0.57,0.69)
Achievements are question of luck	0.43	0.72	0.44
		(0.64,0.79)	(0.3904596 0.4925686)
Other people influence my life	0.55	1.16	0.72
		(1.05,1.26)	(0.65,0.79)
Doubt my abilities	0.55	0.98	0.58
		(0.88,1.08)	(0.53,0.64)
Little control over my life	0.67	1.88	1.33
		(1.60,2.15)	(1.17,1.51)

Table 3.7: Estimates of the Psychometric Question Equations: $M^* = \alpha\theta + \varepsilon_M$

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

	γ	Bayesian confidence intervals
Age	0.04	0.03,0.04
Gender	-0.04	-0.16,0.07
Religion important	-0.04	-0.17,0.10
Low education in 1999	-0.02	-0.16,0.12
High education in 1999	0.17	-0.02,0.35
Father highly educated	-0.01	-0.20,0.18
Mother highly educated	0.34	0.10,0.58
In education in 1999	0.15	0.00,0.30
Turkish immigrant	-1.06	-1.72,-0.39
Central European immigrant	-0.19	-0.76,0.38
EU15 immigrant	-0.60	-1.39,0.16
Turkish second generation	-0.55	-0.86,-0.22
Central European second generation	0.02	-0.37,0.42
EU15 second generation	-0.22	-0.61,0.15
German second generation	-0.12	-0.39,0.15
Immigrant foreign language spoken at home	-0.40	-0.97,0.18
Second generation foreign language spoken at home	-0.10	-0.74,0.53
Time stayed in Germany	0.03	-0.00,0.07

Table 3.8: Estimates of Determinants of the Locus of Control: $\theta = \gamma W + \varepsilon_\theta$, estimated by MCMC

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

In order to see whether the locus of control matters more for immigrants than for natives, the graphs in appendix A show the effect of the locus of control on employment for immigrants and natives and for children of immigrants and natives. The graphs show that a more internal locus of control has a positive for everyone - for immigrants, natives, for children of immigrants and for men and women. They also indicate an apparent gap between immigrants and natives and a smaller gap between children of immigrants and natives.

3.6 Conclusion

This paper set out to examine whether personality traits matter for the labour market performance of immigrants to Germany, using the example of the locus of control - the belief of an individual in their own ability to control their lifecourse. We find that a strong belief in control over one's life has a positive effect on the probability of being employed. Immigrants have a more external locus of control than natives, which means that they believe that their lives are more controlled by external circumstances than by themselves. The second generation also has a more external locus of control than natives, but it is already more internal than that of immigrants. This is evidence for a generational convergence of migrants' locus of control towards that of natives. Immigrants have a double disadvantage on the labour market : they are disadvantaged by lower employment chances because of their status and they are additionally disadvantaged due to having a lower sense of being able to control their life and this sense of control positively matters for the probability of being employed.

There seems to be a barrier in the German labour market towards immigrants, which can be overcome by self-confidence, belief in success, personal dedication and commitment. To a certain degree, an effort to adjust to the new country can be expected of an immigrant. But the German labour market should be of such a structure, that this effort should not be necessary to overcome discrimination, but only to start an adjustment process. German integration policies should include some measures to enhance a stronger belief in immigrants, that they will be able to manage their situation. The paper provides evidence for the success and appropriateness of policies or behavior towards migrants, which encourage their belief in their success.

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

3.7 Appendix: Figures

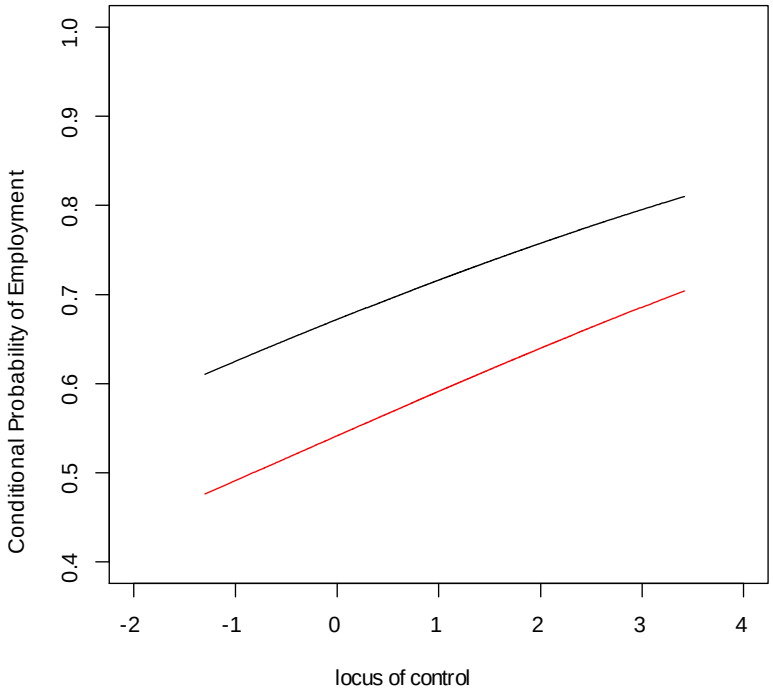

Figure 3.1: Figure 1: The Conditional Probability of Being Employed for Natives vs Immigrants (32 year old, medium educated, married men with children)

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

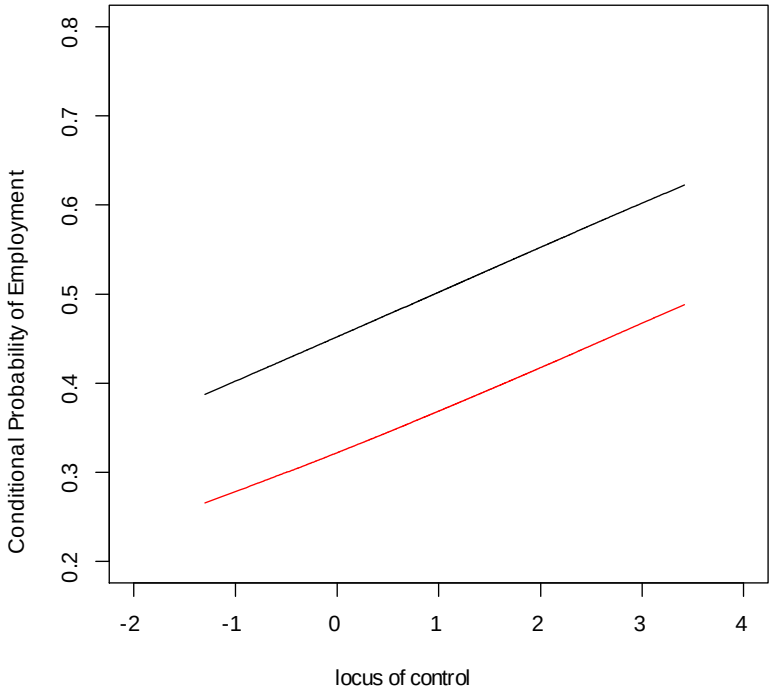

Figure 3.2: The Conditional Probability of Being Employed for Natives vs Immigrants (32 year old, medium educated, married women with children)

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

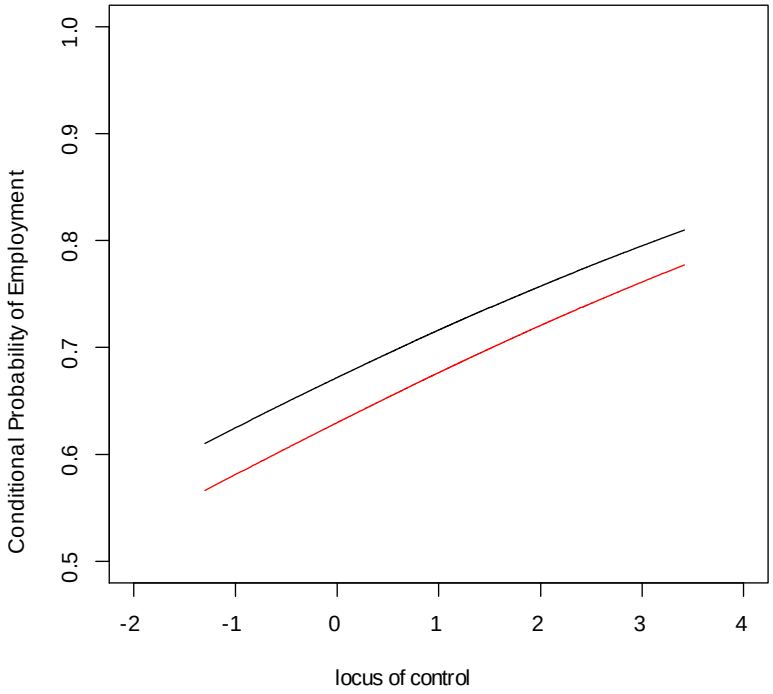

Figure 3.3: The Conditional Probability of Being Employed for Natives vs Second Generation (32 year old, medium educated, married men with children)

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

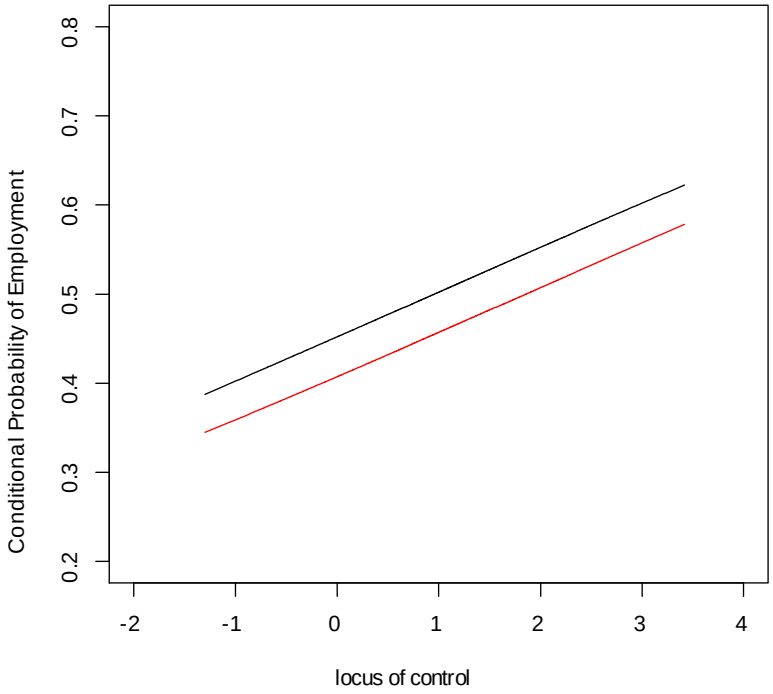

Figure 3.4: The Conditional Probability of Being Employed for Natives vs Second Generation (32 year old, medium educated, married women with children)

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

BIBLIOGRAPHY

[1] Albert, J.H. & Chib, S. (1993) : Bayesian Analysis of Binary and Polychotomous Response Data, *Journal of the American Statistical Association, 88 (422), 669-679.*

[2] Baker, M. & Dwayne B. (1994) : The Performance of Immigrants in the Canadian Labor Market, *Journal of Labor Economics 12.*

[3] Berry, J.W. (2001) : A Psychology of Immigration, *Journal of Social Issues, 29, 541–62.*

[4] Borghans, L.; Duckworth, A.L.; Heckman, J. & terWeel, B. (2008) : The Economics of Psychology and Personality Traits, *Journal of Human Resources, 43, 972-1059.*

[5] Borghans, L.; Golsteyn, B.; Heckman, J. & Meijers, H. (2009) : Gender Differences in Risk Aversion and Ambiguity Aversion, *NBER Working Papers 14713, National Bureau of Economic Research, Inc.*

[6] Borjas, G. (1985) : Assimilation, Changes in Cohort Quality, and the Earnings of Immigrants, *Journal of Labor Economics.*

[7] Borjas, G. (1987) : Self-Selection and the Earnings of Immigrants *American Economic Review* 77.

[8] Bowles, S.; Gintis, H. & Osborne M. (2001a): Incentive-Enhancing Preferences: Personality, Behavior and Earnings, *American Economic Review 91 (2), 155-158.*

[9] Bowles,S.; Gintis, H. & Osborne, M. (2001b): The Determinants of Earnings: A Behavioral Approach, *Journal of Economic Literature 39(4), 1137-1176.*

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

[10] Canerio, P.; Hansen, K. & Heckman J. (2003), Estimating Distributions of Treatment Effects with an Application to the returns of Schooling and Measurement of the Effects of Uncertainty of College Choice, *International Economic Review 44(2), 361-442.*

[11] Chiswick, B.R. (1978) : The Effect of Americanization on Earnings of Foreign-born Young Men, *Journal of Political Economy 86(5), 897-921.*

[12] Chiswick, B.R. (1988) : Differences in Education and Earnings across Racial and Ethnic Groups: Tastes, Discrimination and Investments in Child Quality, *Quarterly Journal of Economics 103 (3).*

[13] Chiswick, B.R. & Miller, P.W. (1999): Language skills and earnings among legalized aliens, *Journal of Population Economics 12(1):63-89.*

[14] Constant, A. & Zimmermann L. & Zimmermann, K.F. (2006): Ethnic Self-Identification of First Generation Migrants, *IZA Discussion Papers*, *Institute for the Study of Labor (IZA).*

[15] Cunha, F. & Heckman, J. (2007) : The Technology of Skill Formation, *American Economic Review 97(2), 31-47.*

[16] Cunha, F.; Heckman, J.; Lochner, L. & Masterov,D. (2006): Interpreting the Evidence on Life Cycle Skill Formation. In Eric A. Hanushek and Frank Welch, eds., *Handbook of the Economics of Education*, chapter 12. Amsterdam, North-Holland, 697–812.

[17] Cunha, F.; Heckman, J. & Schennach, S. (2010) : Estimating the Technology of Cognitive and Noncognitive Skill Formation, *NBER Working Paper No. 15664.*

[18] de Palo, D.; Faini, R. & Venturini, A. (2006) : The Social Assimilation of Immigrants, *IZA Discussion Papers, Institute for the Study of Labor (IZA).*

[19] Duleep, H. & Regrets, M. (1999) : Immigrants and Human Capital Investment, *American Economic Review.*

[20] Duncan, G.J. & Dunifon, R. (1998) : Long-Run Effects of Motivation on Labor Market Success, *Social Psychology Quarterly, pp 33-48* : A Study of Asian Immigrants and their family ties. Kalmazzo MI: Upjohn Institute of Economic Research.

[21] Dustmann, C. (1993) : Earnings Adjustments of Temporary Migrants, *Journal of Population Economics* 6.

[22] Dustmann, C. & Schmidt, C. (2000) : The Wage Performance of Immigrant Women: Full-Time Jobs, Part-Time Jobs, and the Role of Selection, *IZA Discussion Papers* 233, *Institute for the Study of Labor (IZA).*

[23] Eeckhout, J. & Weng, X. (2009) : Assortative Learning, *working paper*

[24] Eckstein, Z. & Weiss, Y. (2002) : The Integration of Immigrants from the Former Soviet Union in the Israeli Labor Market", in Ben-Bassat Avi, (ed.), *The Israeli Economy, 1985-1998: From Government Intervention to Market Economics,* Essays in Memory of Prof. Michael Bruno, MIT Press. 2002, pp. 349-378.

[25] Fahrmeir, L. & Raach, A. (2006) : A Bayesian semiparametric latent variable model for mixed responses, *Psychometrika.*

[26] Farkas,G. (2003): Cognitive Skills and Noncognitive Traits and Behaviors in Stratification Processes, *Annual Review of Sociology, 29, 541–62.*

[27] Fertig, M. & Schmidt, C. (2001) : First- and Second-Generation Migrants in Germany, What do we Know and What do People Think, *IZA Discussion Papers, Institute for the Study of Labor (IZA).*

[28] Fertig, M. (2004): The Societal Integration of Immigrants in Germany, *IZA Discussion Papers, Institute for the Study of Labor (IZA).*

[29] Gang, I. & Zimmermann, K.F. (1999) : Is child like parent? Educational Attainment and Ethnic Origin, *Journal of Human Resources.*

[30] Heckman, J. (1995) : Lessons from the Bell Curve, *Journal of Political Economy 103 (5), 1091.*

[31] Heckman, J.; Stixrud, J. & Urzua, S. (2006) : The Effects of Cognitive and Non-cognitive Abilities on Labor Market Outcomes and Social Behavior, web supplement available at jenni.uchicago.edu/noncog, *Journal of Labor Economics.*

[32] Herrnstein, R.J. & Murray, C.A. (1994) : The Bell Curve: Intelligence and Class, *Structure in American Life, New York, Free Press.*

[33] Hinte, H. & Zimmermann, K.F. (2005) : Arbeitsmarkt & Zuwanderung, Springer, Berlin, Heidelberg.

[34] Kahoe, R. (1974): Personality and achievement correlates of intrinsic and extrinsic religious orientations, *Journal of Personality and Social Psychology 29: 812–8.*

[35] LaLonde, R.H. & Topel, R.H. (1992) : Immigrants in the American Labor Market : Quality, Assimilation and Distributional Effects, *American Economic Review* 8.

[36] Liebig, T. (2007) : The Labor Market Integration of Immigrants, *OECD Social Employment and Migration Working Papers.*

[37] Matzkin, R. (2003) : Unobservable Instruments, *mimeo*, Northwestern University.

[38] Matzkin, R. (2007) : Nonparametric Identification, *Handbook of Econometrics Vol 6B.*

[39] Mueller, G. & Plug, E. (2006) : Estimating the Effect of Personality on Male and Female Earnings, *Industrial and Labor Relations Review 60 (1), 3-22.*

[40] Osborne, M. (2000): The power of personality: Labor market rewards and the treansmission of earnings, *Electronic Doctoral Dissertations for UMass Amherst, http://scholarworks.umass.edu/dissertations/AAI9988828.*

[41] Osborne-Groves, M. (2006) : How important is your personality? Labor market returns to personality for women in the U.S. and U.K., *Journal of Economic Psychology.*

[42] Rabe-Hesketh, S. & Skondral, A. (2004) : Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models. Boca Raton, FL: Chapman & Hall/CRC.

[43] Robert, C.P. & Casella, G. (2004) : Monte Carlo statistical methods (2nd edition). Springer, New York.

[44] Rotter, J.B. (1966) : Generalized Expectancies for Internal versus External Control of Reinforcement, *American Psychological Association.*

[45] Spearman, C. (1904) : General Intelligence, Objectively Determined and Measured, *American Journal of Psychology 15, 201-293.*

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

CHAPTER 4

ETHNIC IDENTITY AND EDUCATIONAL OUTCOMES OF GERMAN IMMIGRANTS AND THEIR CHILDREN

4.1 Introduction

Education is a crucial factor to integrate youth with an immigrant background into their host society. This is especially the case for Germany where firstly, a high importance is attached to formal educational degrees, and secondly, there is a high need for skilled labor – of individuals with vocational as well as tertiary educational[1] backgrounds. Education also plays an important role in the political strategy of the European Union for the next twenty years. One of the criteria that each the EU countries should achieve by 2020 is that 40% of the 30-34 year old population should have tertiary education[2]. In Germany, as in many other European countries, immigrants and their offspring have on average lower educational attainment levels than natives and their children and it could of political interest to increase educational attainment of immigrants and their children to meet the Europe 2020 criterion.

In this paper we aim to estimate the effect of a measure of immigrants' and their children's identification with German society on educational choices. Not feeling part of society can provoke the creation of a concept known in sociology as "oppositional identities" - individuals identify themselves with values opposite to those of the majority. For example, if the native population is characterized by a good education, the immigrant population would not strive for a good education in order to identify themselves as the opposite of the native population.

Do educational outcomes differ for individuals of different ethnic backgrounds? In the economic literature there is theoretical as well as empirical evidence that for educational outcomes, ethnicity matters. Prominent examples for research on the case of the United States are Borjas (1992), Chiswick (1988), Chiswick & Miller (1994) and Duleep & Regrets (1999). For the German case, the field is still in development, but main contributions have been made by Fertig (2001) Gang, Zimmerman (1999) and Riphahn (2001). Chiswick, DebBurman (2003) state that differences in educational attainment persist over immigrant

[1] Tertiary education is the education following the completed education in a school providing secondary education, such as high school, secondary school, gymnasium or university preparation school. It includes undergraduate and graduate studies or vocational training.

[2] See "Europe 2020" by the European Commisssion, 2010.

generations.

Which aspect of ethnicity causes the difference? In this paper we hypothesize, that the immigrant's identification with the German society could cause a lack of educational performance in the host country society. Akerlof & Kranton (2002) have reviewed the literature in other academic fields on the role of identity for schooling. Identity is seen as a driving force for educational success in this essay. If this is the case, then ethnic identity is likely to affect educational outcomes of immigrants and their children in Germany.

We conceptualize ethnic identity as a latent factor which is made manifest by a set of ethnic identity questions posed to immigrants and their children in the German Socioeconomic Panel (GSOEP) in the waves 1999-2001. In the relevant sociological and anthropological literature there is a debate as to whether ethnic identity of immigrants is to be seen as a one-dimensional "either one culture or the other" or as a two-dimensional concept signifying an identification with both cultures. Since we focus on the problem of endogeneity of the measure in this paper, we choose to examine a one-dimensional concept.

This question of educational outcomes of immigrants and its determinants gains current relevance because the German government initiated in 2008 a nation-wide qualification initiative "Aufstieg durch Bildung" (progress by education). It is furthermore a crucial question since every fifth adolescent with immigrant background drops out of school without any degree. Additionally they are under-represented among university students and upper secondary scholars . Fertig & Schmidt (2001), show that second generation migrants perform worse in educational attainment than natives or first generation migrants in the Mikrocensus 1995. This "dissimilation" across generations of immigrants in Germany is also shown in work by Riphahn (2000). These facts advocate a public initiative in education policy directed exclusively towards immigrants.

This paper finds that identification with German culture positively affects the educational outcomes of immigrants and their children. Second generation immigrants identify themselves more strongly with the German culture than their parents' generation.

The second section analyzes previous literature on identity and education, identity in economics and ethnic identity in economics is analyzed. The third section explains the empirical strategy and describes the sample. In the fourth section the results are analyzed and the fifth section concludes.

4.2 Theory and Previous Literature

A rather recent body of literature in economics considers educational assimilation of immigrants. The study of assimilation of immigrants in economics began with examining

immigrants' socioeconomic success such as earnings or employment. Educational attainment of immigrants is analyzed later and is based on previous economic human capital theories. The role ethnic identity or other psychological factors play in the determination of these measures is acknowledged by some economists and is connected to previous theoretical studies of the role of identity in economic returns. Culture is considered a relevant issue in the migration literature but has not yet been widely analysed in the economic literature (Epstein, Gang 2010).

4.2.1 Ethnicity and Socioeconomic Success

Cameron and Heckman (1999) studied the determinants of educational attainment of different ethnic groups in the US. They employed a dynamic discrete choice model with an underlying one-factor structure and showed differences across ethnicities in change of schooling decisions in response to rising returns to schooling, parental income, parental background, tuition rates and county specific variables. They emphasize the importance of the role of long-term parental background rather than income.

Gang and Zimmermann (1999) estimated the effect of ethnic origin on educational attainment for German second generation migrants. They found that the size of the ethnic network matters as well as the ethnic origin for educational attainment of second generation migrants. Parental education does not matter.

The authors mention two problems. Firstly, the measure of ethnic origin is the passport[3]. It is possible that while an individual might hold a German passport he or she is nevertheless foreign. And secondly, they suggest a problem of endogeneity between ethnicity and educational attainment. I aim to address both these problems, by approximating ethnicity by a latent factor.

Measuring ethnic identity of German first generation migrants has been done by Constant and Zimmermann (2006, 2007), Constant, Gataullina and Zimmermann (2006,2007), Gataullina, Zimmermann (2006) and Bonin et al. (2006). The authors construct a measure of ethnic identity based on two axes : identification with the host country and with the country of origin. They found, that an individual related strongly to both cultures, is predicted to be more successful on the labor market.

[3]See page 558.

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

4.2.2 The Role of Educational Attainment in Economics

The theory of human capital investment developed by Becker (1964) is a main building block for economists to understand the economic meaning of educational attainment. It states that an individual invests in human capital to maximize net wealth. Chiswick (1978) has extended this model to account for intergenerational differences in educational attainment. He has additionally formalized the concept of "international transferability of skills".

Economics has so far focussed on assessing the role education resources play for the educational attainment of pupils, less so the individual's characteristics.

4.2.3 Determinants of Immigrants' Educational Attainment

Chiswick and DebBurman (2003) stress, that educational attainment of immigrants needs to be studied separately for each generation of immigrants. The reason is the difference in where the education was received. Such a differentiating analysis allows the authors to separate an intergenerational transmission effect of educational attainment from the host country society's effects.

Among the main determinants Chiswick & DebBurman name country of origin and age at immigration or whether the immigrant was born in the host country. That is, generation and ethnicity matter.

Which aspect of ethnicity matters? Is it nationality, in that some cultures perform higher than others or some cultures are closer to the host country and can therefore adapt to the host country education system more easily? Or is it the assimilation of an individual that makes him more motivated?

In this paper, we focus on the latter. As Akerlof and Kranton (2002) state, we postulate that it is crucial for an immigrant's educational attainment to identify with the host country. We will explain why the inclusion of a concept of identity could change economic findings for the differences in educational attainment between immigrants and natives.

4.2.4 A Theory of Ethnic Identity

Ethnic Identity is a concept, which we distinguish from ethnicity. The latter can be identified by an individual's passport or nationality. The former refers to a socio-psychological, cultural-psychological and even anthropological phenomenon. In non-economic sciences, the necessity of this distinction has been acknowledged. Constant and Zimmermann (2006) and Constant, Gataullina and Zimmermann (2006) make this distinction as economic authors.

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

We follow their example.

A definition of ethnic identity can be found in Phinney (1992)[4], who is seen as a major contributor to the literature of ethnic identity in psychology (see Worrell et al (2006)) : ethnic identity consists of "a feeling of belonging to one's group, a clear understanding of the meaning of one's [group] membership, positive attitudes towards the group, familiarity with its history and culture, and involvement in its practices".

Worell (2006) argues that concepts of this type apply mostly in an ethnically diverse society. Phinney argues even further that a concept of ethnic identity is necessary only in pluralistic societies. It has the function of giving individuals a way to secure their identity towards one, that builds on different principles. ("the concept of ethnic identity provides a way to understanding the need to assert oneself in the face of threats to one's identity", Phinney (10992) p. 499). To understand ethnic identity, it is helpful to consider the context.

A well established conceptualization of an immigrant's ethnic identity,developed by Berry (1980) et al, is to define the categories "assimilation", "integration" , "separation", "marginalization". It is a nonlinear theory, allowing for a two-dimensional concept of ethnic identity : an axis for a connection to the home country and an axis for the connection to the host country.

Two conflicting theories argue for and against orthogonality of identification to different cultures. If an individual feels connected to one culture, does it mean he cannot feel connected to another or can he be connected to both. The first case would be the belief underlying the linear theory and the latter a non-linear theory. The linear theory imposes an axis with both cultures at each end of the spectrum, while the non-linear theory allows for two axes, one for each culture.

In this paper we focus on the one-dimensional ethnic identity concept.

4.2.5 Identity

Noneconomic disciplines have addressed the study of identity earlier than economics. The developmental psychologist Erik Erikson worked on identity and personality. Erikson (1950, 1959) states the adolescent's identity crisis and has started a wide field research on this topic in psychology. In psychology identity is seen to positively affect psychological indicators such as self-esteem. Erikson defines identity in a social context, which is also done by other scholars of the field, in opposition to scholars studying the ego side of identity. The belief in a social side of identity leads to studies of social identities, such as racial identity.

In economics, Akerlof and Kranton (2000) have constructed a micro-economic model,

[4]See page 169.

promoting the introduction of identity into the utility function. As in Erikson's approach, identity is modelled in terms of belonging to a group and following the prescriptions of this group. Akerlof and Kranton (2010) gives the most recent overview of the research on the role of identity in economics. Can identity effect outcomes, another economic indicator?

4.2.6 The Role of Ethnic Identity in Education

Akerlof and Kranton (2002) state that identity affects the amount of effort an individual devotes to education[5]. The students can decide whether or not to adapt a school's social category, which contains an image of a certain effort level. Can ethnic identity affect this sorting? The authors have reviewed literature in other academic fields. Historians see the choice of an individual to adapt to a school's social category as problematic, when the individual's background conflicts with the school's ideal. They identify a clash arising from an "Americanization" of schools in the early twentieth century on one hand and a rise of the proportion of immigrants in American schools on the other hand. A school can be seen as representing home country ideals, and therefore, an individual with a foreign background, might experience this conflict.

Sadowksi (2001 in Akerlof and Kranton (2002)) states that a large amount of the gap in test scores between African-Americans and Americans is accounted for by "a feeling of connectedness" to the host-country (in this case white) schools.

The hypothesis of this paper is that this conflict between background and host country society could cause differences in educational attainment unaccounted for by nationality differences. Akerlof and Kranton (2000) mention this sociological phenomenon briefly[6]. Noneconomic literature states, that such a conflict can in the worst case cause rejection of the school. As stated above,this is the case for a large part of youth with migration background in Germany.

4.3 Empirical Strategy

4.3.1 The Model: Generalized Simultaneous Equation System

The model is a generalized simultaneous equation model for the educational outcomes of German immigrants and their children and for the measures of their ethnic identification. Such a model can be viewed as a factor model embedded in a simultaneous equation model. This way of proceeding is close to LISREL models and models analyzed in AMOS. We

[5] See page 1168.
[6] See page 1171.

estimate using a methodology in the spirit of common factor analysis in structural equation modelling. This technique is based on work by Heckman, Stixrud and Urzua (2006) and Fahrmeier and Raach (2006). An identification strategy is provided in Carneiro, Hansen and Heckman (2003).

Underlying the estimation procedure is a confirmatory factor analysis. In contrast to an exploratory analysis this means that we presuppose the number of factors underlying the model and estimate simultaneously coefficients on observables, factor loadings and factor scores, which are the realizations of the latent variable for each individual. These values are obviously subject to the assumptions made and therefore the data fitting process.

We estimate using a Markov Chain Monte Carlo (MCMC)[7] method, based on work by Heckman, Stixrud and Urzua (2006), Carneiro, Hansen and Heckman (2003) and Fahrmeir and Raach (2006). A thorough explanation and discussion of the methodology and simulation results are given in chapter two of my thesis. This methodology allows us to analyze a small sample, receive estimates of the realizations of the latent concept for each individual simultaneously with the other estimates of the model and estimates of posterior standard errors. This method is an alternative to a maximum likelihood procedure.

The advantages of simultaneous estimation are :

- Measurement error of the items is taken into account
- Endogeneity between the ethnic questions and educational attainment is addressed by introducing a factor structure which controls for the interdependence. Therefore the error term is not correlated with independent variables of the model
- Efficiency : we use more information on the latent concept.
- Interpretability : embedding a factor model in a SEM framework reduces subjectivity inherent to factor analysis

The main disadvantage is, that a possible miss-specification in the measurement equations can enter the outcome model and bias the remaining estimates. Therefore we need to argue that the assumptions required for consistent estimates are verified by the data.

4.3.1.1 The Setting

An important challenge of our analysis is to address the problem of endogeneity of ethnic identity in an educational outcome equation. We address this problem by taking

[7] In Appendix A we show the algorithm used for the estimation.

a measure of ethnic identity in 1999 and a measure of educational outcome in 2007. We capture and examine the effect that ethnic identity acquired in 1999 has on education in the years 1999-2007. Our methodology allows us to account for determinants previous of 1999 of ethnic identity in 1999, such as parental background, time since immigration (for the first generation) and ethnic background. The idea is to capture a part of the cycle of education that forms identity and identity that determines educational choice. So we measure identity in 1999, when individuals in the sample are aged 17-32 and control for their education level at that point by including education as a determinant of θ. We then capture the effect of θ on education 8 years later in 2007. This procedure accounts for the fact that ethnic identity is determined by previous education and addresses the problem of reverse causality.

A possible question is, whether the age group of 17-32 has not already finished their education and would not change it after 1999. As we show in the descriptive section below, this is not the case and educational outcomes still change after 1999, for all ages between 17 and 32. In Germany this is likely to be the case, firstly because there are 13 years of schooling and men need to add an additional social or military year before they can start studying. Secondly, university education took a long time before the introduction of the bachelor/master system. Thirdly, there is a well-established system called "zweiter Bildungsweg", which means translated freely "second chance education". This system allows people living in Germany to take another chance in evening schools to gain higher educational levels. Fourthly, there are the so-called "Berufsakademien" (professional academies) in which young people obtain a higher educational level mainly in vocational training and can work at the same time. This system allows to earn money, professional experience and education at the same time.

The model can be divided into two parts, an outcome model for the educational outcome and a measurement model for the psychometric questions. It is a parametric linear additive model. This is on the one hand a considerably restrictive setting and on the other hand convenient to specify latent variables and their effect in a consistent (but restrictive) way. The model is a set of simultaneous probit-structure models with an endogenized latent factor.

$$\begin{aligned}
Y_{07i} &= k_Y \ if \ c_{Y-1} < Y^*_{07i} < c_Y \\
Y^*_{07i} &= \beta^Y X^Y_{07i} + \alpha^Y \theta_{99i} + \varepsilon^Y_i \\
M_{99i} &= k_M \ if \ c_{M-1} < M^*_{99i} < c_M \\
M^*_{99i} &= \alpha^M \theta_{99i} + \varepsilon^M_i \\
\theta_{99i} &= \gamma W_{99i} + \varepsilon^\theta_i
\end{aligned}$$

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

Y_{07i} is a scalar tricategorical ordered variable and denotes the educational outcome in 2007. Y^*_{07i} is a scalar denoting the latent underlying variable of the ordered response Y_{07i}. We denote the tricategorical responses to the psychometric questions in 1999-2001 with M_{99i}.M_{99i} is a vector containing a set of the psychometric questions. M^*_{99i} is equally a vector of the size of the number of psychometric questions and denotes the latent underlying variables or thresholds for the ordered responses. X^Y_{07i} denotes the observable determinants of the educational outcome Y_{07i}. θ_{99i} denotes a latent factor. The latent factor is endogenized and therefore is specified as a dependent variable on the observable determinants W_{99i}.γ denotes the set of coefficients of the vector W_{99i}. The vector α^Y and α^M denote the factor loadings. Factor loadings can be interpreted as the correlation of the latent factor with the dependent variable. c_Y and c_M denote cutpoints for the ordered responses. β^Y denotes the set of coefficients for the observable determinants of Y_{07i}. ε^Y_i,ε^M and ε^θ_i are normal random error terms.

4.3.1.2 Assumptions

To be able to identify the model, we make the following independence and distributional assumptions as well as some normalizations and assumptions common for factor analytical models. These assumptions are discussed and explained in detail in section 2.2.1.1. At this point we mention the assumptions - adapted to the model in this chapter - and refer to section 2.2.1.1. for their detailed discussion. For simplicity we suppress the time and individual subscript in the following.

First we make an assumption typical in factor analysis and ensuring the uniqueness of the factor and its loadings.

$$Var(\theta) = 1$$

As shown in section 2.2.1.1. the latter assumption, the model and a distributional assumption imposed on ε^θ imply that

$$\theta \sim N(\gamma W, 1)$$

Next we impose several distributional assumptions

$$\begin{aligned} \varepsilon^\theta &\sim N(0,1) \\ \varepsilon^M &\sim N(0,1) \\ \varepsilon^D &\sim N(0,1) \end{aligned}$$

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

In addition we make the following (conditional) independence assumptions

$$\begin{aligned} W &\perp \varepsilon^{\theta} \\ \theta &\perp \varepsilon^{M}|W \\ \theta &\perp \varepsilon^{Y}|W \\ \theta &\perp X^{Y}|W \\ X &\perp \varepsilon^{Y} \end{aligned}$$

We also assume the following local independence conditions (see for example Rabe-Hesketh 2004), which are typical in latent variable modelling:

$$\begin{aligned} Y &\perp M|\theta \\ M_i &\perp M_j|\theta \quad \forall i \neq j \end{aligned}$$

Finally we impose normalizations on the cutpoints between the first and the second category of the tricategorical items and outcome variable.

$$\begin{aligned} c_1^Y &= 0 \\ c_{1j}^M &= 0 \ \forall j \end{aligned}$$

4.3.1.3 Discussion of Conditions and Interpretation of the Latent Factors

The model and its conditions imply, that once we know the observable control variables and the latent factors, educational outcome is independent of the psychometric questions. We control for parental background, childhood characteristics, immigrant generation and ethnic background in addition to the ethnic identity factors. Both the conditional independence and the independence conditions, in addition to the choice of psychometric questions, serve to interpret the latent factors. The latent factors are latent and therefore need to be interpreted. We know that they are a variation, which is informative for the dependent variable. They are by construction independent of specified observable variables. These facts restrict them to signify a specific "variation".

4.3.2 Measuring Ethnic Identity : Psychometrics

Measurements of ethnic identity are of concern in cross-cultural psychology and in social psychology. For quantitative analysis in psychology, one methodology is psychometrics: the design of questionnaires to extract a common latent factor underlying a response pattern.

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

Alternatively the questions could be treated as proxies for the latent concept and therefore as conventional explanatory variables. A problem with this approach is the interpretability of the coefficients on the questions. Factor analysis helps to identify a concept or factor of interest underlying a related set of questions. The coefficient on this factor – the factor loading – can then clearly be interpreted as the amount of variation in the data, that is explained by the latent factor.

In this study we aim to estimate the effect of a complex concept – ethnic identity. We believe that this concept cannot be proxied by a single variable. Factor analysis and a suitable choice of items allows to cover several dimensions of the concept "ethnic identity". To use the appropriate psychometric measures, we need to choose questions that could capture a latent factor called "ethnic identity" using factor analysis.

There are numerous psychometric studies proposing sets of such questions (items). Phinney (1992) constructed a set of 20 questions called "multigroup ethnic identity measure" MIEM especially for adolescents. It is a set of items most frequently used to study ethnic identity (see Worrell 2006, p.38). The same questions can be used across different ethnic groups. Phinney's factor analysis shows that two interpretable factors can be extracted : a concept, which could be identified as ethnic identity search (EI), and one of commitment/belonging to another group (OGO). This scale is matter of discussion in terms of structural validity - does it really measure a concept that could be called "ethnic identity".

As mentioned above, we use a linear one-dimensional measure of ethnic identity psychometric ethnic identity studies promote non-linear theories of ethnic identity, allowing for a factor, or an axis, for each culture. These axes need to be orthogonal to each other, to satisfy a crucial assumption in standard factor analysis. Concerning the development of an index containing an identification with two cultures, for example the host country and the country of origin, Phinney (1992) notes, that attitudes towards other groups than their own, should be seen as "a factor ".

In the spirit of the analysis in psychology, a small body of literature in economics has addressed the problem of empirical measurement of some concept of ethnic identity by survey questions. Nekby and Roedlin (2007) use questions concerning language spoken, language skills and ethnic activities. This is in line with Constant and Zimmermann (2006), who state main elements of ethnic identity : language, visible cultural elements. These authors add ethnic self-identification, ethnic networks and citizenship plans. In the following I delineate the process we propose for selection of questions for this study. We compare estimates using a one-dimensional as well as a two-dimensional concept.

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

4.3.3 Measuring ethnic identity : selecting the items for a one-dimensional identity concept

We need to select a set of questions, that allows a convincing analysis of the concept "ethnic identity". First of all, the questions selected for this study need to cover main dimensions of ethnic identity. We follow Constant and Zimmermann (2006), who name ethnic self-identification, involvement in visible ethnic activities, language use, belonging to a social network (ethnic interaction) and migration history/citizenship plans. At first we adopt this categorization. We select the number of items per dimension making sure not to suboptimize, that is we avoid adding so many variables for one dimension that a new factor can be identified.

In a pre-analysis we analyze solely the measures (items) and analyze the correlation matrix. We choose a set of those items that have bivariate correlations above 0.3. This is the customary threshold in settings of this type. We find that seven questions satisfy this condition.

With the set of questions in hand we then proceed to a confirmatory factor analysis embedded in a structural equation model. That is, we add an a priori theory in form of explanatory variables, outcome variables (educational attainment) and the number of factors in the model. The confirmatory analysis has the purpose to test this imposed a priori theory, while the exploratory analysis had no underlying presumptions about the latent structure underlying the measures.

4.3.3.1 One-dimensional Model

We found seven items that yield one component for our one-dimensional factor analysis. The items reflect the dimensions of "ethnic identity" Constant and Zimmermann (2006) have identified : ethnic self-identification, visible cultural activities, social network/ethnic interactions, language.

All questions are tricategorical ordered responses. Possible responses are indicated after naming the question.

- In your opinion, how well do you speak German? 1 - fairly,poorly, not at all, 2 - good, 3 - very well

- In your opinion, how well do you write German? 1 - fairly,poorly, not at all, 2 - good, 3 - very well

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

- Which language do you use at home? 1 - mostly language of country of origin, 2 - both equally, 3 -mostly German
- To what extent do you feel German? 1 - barely or not at all, 2 - in some respects, 3 - completely or mostly
- When you read the newspaper: do you read newspapers from... 1-only or mostly host country, 2-from both or not at all,3-only or mostly country of origin
- How often do you cook meals traditional to your country of origin? 1-only or mostly host country, 2-from both or not at all,3-only or mostly country of origin
- Of which origin are your three best friends? (constructed variable)1 - three friends country of origin , 2 - mixed friends, 3 - three friends German

Our reasoning, why we call this concept "ethnic identity" is, that it covers a range of dimensions constituting ethnic identity. If we choose to believe in a one-dimensional ethnic identity concept, we argue to employ in the Generalized Simultaneous Equation Model the questions mentioned above for the following two reasons : (1) statistically this set of data contains a unique component, which can explain the total variance of the data (the common and the unique variance) and (2) it reflects a set of dimensions, that can be theoretically shown to account for a concept of ethnic identity.

4.3.4 Educational Outcomes: The German Education System

Several studies seek to explain completed years of schooling. Others consider completed schooling degrees or the probability to go to college. Cameron and Heckman (1999) claim, that these studies mask the different factors at play determining each stage. If sufficient longitudinal data is at hand, a dynamic model of educational attainment, in the spirit of Cameron and Heckman (1999) is suitable.

The German education system consists of three stages. The first stage consists of 4 years of primary school, which every individual needs to complete. The second stage consists of three choices : secondary school, intermediate school, and upper secondary school. These last for five,seven and nine years. There is a small number of integrated schools and schools for disadvantaged children, but they have not reached the level of being a main type of school. It is mandatory to have at least a secondary degree.

The possibilities to choose a tertiary level school depend on the type of secondary schooling. Tertiary level choices are "university/technical college (Fachhochshule)", "vocational

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

school (Berufsschule)" or "apprenticeship (Lehre)". Secondary and intermediate schools allow completion of an apprenticeship and vocational schooling, even though it is easier to get a place in a vocational school with an intermediate school degree. Upper secondary schooling allows entering university or technical college.

There is the small possibility to enter the "second education path", which allows to catch up on a certain educational level, such as passing an upper secondary school degree at an evening school.

To construct our tricategorical ordered measure of educational outcome can take the values "low", "medium" and "high". We use the ISCED[8] standardization code: A low education level comprises pre-primary education, primary education or first stage of basic education and lower secondary or second stage of basic education (ISCED 0,1,2). A medium education level is classified as (upper) secondary education or post-secondary non-tertiary education (ISCED 3,4). A high education level is the first stage of tertiary education or the second stage of tertiary education (ISCED 5,6).

4.3.5 Sample Description and Variable Definitions

We use the 2007 wave of the German Socioeconomic Panel (GSOEP) as well as the 1999, 2000 and 2001 waves. The GSOEP is a longitudinal micro dataset for 1984-2007, in which both German citizens and migrants are analyzed. The sub-sample of the GSOEP analyzed in this paper consists of 540 individuals, who were born in Germany and do not have German nationality at birth and those who were born abroad and who have a foreign citizenship at birth. Unfortunately, we needed to drop individuals with German nationality at birth since these individuals are asked to skip the ethnic questions in the questionnaire. This makes it impossible to conduct a comparison between immigrants, second generation immigrants and those of German origin. About half the sample are male and half are female. We divided the ethnic origins into three major geopolitical groups. The first group holds an EU15, Swiss or US citizenship, making up 33.4% of the sample. The largest groups within this group are Italians and Greek. The second group are those of Turkish origin, making up 29.8% of the sample. The third group consists of those holding a citizenship of Central Europe or of the former Soviet Union. This group accounts for 21.4% of the sample. 15.3% hold German nationality. They are aged from 25-40 in 2007. This means that the individuals are aged 17-32 in 1999.

Immigrants are defined as *foreign-born with no German nationality at birth* and second

[8]See UNESCO (2006) : ISCED 1997 - International Standard Classification of Education, *www.uis.unesco.org.*

generation immigrants are defined as *German born with no German nationality at birth.* Due to a change in German citizenship law in 2000, after which children of immigrants born in Germany after 2000 receive German citizenship and children of immigrants born before 2000 can acquire German citizenship more easily (ius terrae), some of the second generation immigrants in our sample have acquired German citizenship.

4.4 Results

To begin the analysis of the results we study some descriptive statistics on the relationship between educational attainment, identity, immigrant generation and age. First of all we would like to see whether there is a positive relationship between identity and educational attainment over all age groups and both immigrant generations. Secondly, we are interested in the change of educational attainment from 1999 to 2007. Our sample population is 17-32 in 1999 and 25-40 in 2007. In Germany it takes on average longer than in other countries to achieve tertiary education . High school degrees are obtained after 13 years of school and for men there is an obligatory military or civil service lasting at least ten months. These facts push the university starting age to the early 20ies. Before the bachelor/master system was adopted in Germany, it took on average about 6 years to obtain a university degree. The educational system in Germany is quite flexible in the sense that it is technically possible reach a tertiary degree from any secondary level by taking evening classes to obtain a baccalaureate. Another specificity of the German system is the so-called "dual system". This system expresses the importance attributed to technical studies. One can reach a tertiary level in Germany by being a foreman for example. This fact is embodied in the educational variable we chose in the German Socioeconomic Panel.

4.4.1 Descriptive Results

First we study the distributions of educational levels in 1999 and in 2007 by age group, identity percentile and immigrant generation versus native Germans presented in the tables in the Appendix. The tables show absolute and relative frequencies for each educational outcome[9]: "in school" signifies the group of individuals who are in vocational training or who have not yet reached any degree yet. The minimum schooling in Germany is the tenth grade or the degree of the lowest level school ("Hauptschulabschluss"). The category "low"

[9]For a description of the ISCED categorization in the German Socioeconomic Panel see Fuchs and Sixt (2008), page 11.

comprises individuals with an ISCED level of 1,2 or 3[10]. This means the individual has either a degree from the "Hauptschule" or from the next highest level "Realschule" or does not have any degree (and is momentarily not in school to reach a degree). The next highest level in our classification is "medium". Individuals in this class have obtained a degree that corresponds to the category ISCED 3 or 4. In the German system this means that the individuals have either obtained a degree to be able to go to university - the "Abitur", a degree of a specialized high school ("Fachoberschulabschluss") or vocational degree. The category "high" includes individuals with a civil servant education ("Beamtenausbildung"), a degree as a foreman ("Meister", "Techniker") or a university degree.

For the group of 17-20 year old population we find that the overall distribution of educational attainment levels in 1999 is similar to those of the natives in the same year. Although there are only few observations in some of the categories, we can see that individuals in higher identity percentiles tend to have higher educational outcomes. We can see that the distribution over the schooling levels changes between 1999 and 2007 and logically the latter distribution stochastically dominates the former for both second generation immigrants and for natives. Individuals tend to increase their educational attainment levels over time. The 1999 picture is emphasized in 2007 - the higher the identity percentile the higher the education level tends to be. In 2007, the native statistically dominate the second generation. This is probably the case because those who were in the category "in school" in 1999 are in 2007 divided among the categories "low","medium" and "high" and immigrants and their children tend to be more strongly present in lower categories than natives.

This picture is repeated in the category of the 21-26 year old population. The distributions of educational outcomes for second generation immigrants, immigrants and natives change between 1999 and 2007 - the distributions in 2007 stochastically dominate those of 1999 meaning that individuals tend to ameliorate their education as in the group of 17-20 year old population. For the group of 21-26 year old population for both immigrants and the second generation there is clear evidence of the distributions for higher identity percentiles stochastically dominating those for the lower identity percentiles.

27-32 year old individuals also tend to ameliorate their education. In this category no one is in the category "in school". As in the younger categories we can see that the distributions in 2007 stochastically dominate those of 1999.Those of the natives stochastically dominate those of the second generation which dominate those of the immigrants. Being in a higher identity percentile increases the probability of having a higher educational attainment level.

[10]The ISCED levels are labelled differently in the GSOEP than the labels given by the UNESCO and the OECD, but they correspond to the the logik of the ISCED 1997 classification.

Overall the tables show that individuals tend to be in higher educational levels in 2007, that is they still change their educational levels even after 1999 at all age groups. They also show that individuals in a higher percentile of German identity tend to have higher educational attainment levels. This is the case for immigrants and natives. This finding is stronger in 2007 than in 1999 for both immigrants and natives. For all age groups and identity quantiles the distributions over education in 2007 stochastically dominate those in 1999.

4.4.2 Regressions

Before introducing the German identity concept into the model we test a basic model of education and immigrant generation controlling for age and gender, shown in table 4.1. We find, as expected, that the second generation has a significantly higher probability of being in the highest education category, with a coefficient of 0.38. Women have significantly lower educational levels and older immigrants and their children seem slightly less well educated, but this effect is not significant. In a next step we introduce German identity into the model and present results for the one-dimensional endogenous ethnic identity model. We compute the highest posterior density intervals out of the 5%-95% empirical quantiles and verify whether it includes 0. If the Bayesian confidence interval includes 0 the parameter is not significant at the 5% level in the frequentist sense.

Table 4.1 shows the estimates of the educational outcome equation in 2007. A stronger identification with German society has a positive and significant effect on educational outcome in 2007. When controlling for identity age has a positive effect and the coefficient for being female has a negative sign and is significant. The regression includes also control variables for different nationalities in 2007, interacted with the immigrant generation dummy. The base category is "EU15 immigrant". We include nationality since educational attainment might differ across ethnic groups in Germany. Turkish immigrants seem on average less well educated than EU15 immigrants - results show that Turkish immigrants have a lower probability of being in the highest educational outcome category than EU15 immigrants. Central European immigrants have a higher probability. A Turkish immigrant with an identity measure, which is by one-sigma higher than that of an EU15 immigrant would have the same probabilities for the different educational attainment levels as an EU15 immigrant. For the second generation the difference between Turkish and central European origin are not significant as both groups show a coefficient of about -0.3. EU15 second generation citizens seem to have a slightly higher educational level than the first generation EU15 citizens. The difference between the first and the second generation in educational outcomes is no longer

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

significant once we add the German identity measure to the estimated model. We include ethnic background variables, immigrant generation, age and gender both in the educational equation and in the ethnic identity equation both because we believe and are interested in an effect of these variables on ethnic identity and in order to control for any possible correlation between the determinants X of education and θ.

In table 4.2 we present the estimated factor loadings. They represent the effect of the latent factor on the psychometric items. All estimates are positive and significant.

Table 4.3 shows the estimates of the determinants of ethnic identity. We include in the determinants of θ some additional controls, which are not included in the set of determinants of education. The reason for including education is - as outlined above - to address a potential reverse causality problem between education and identity. We include parental education only in the ethnic identity equation since we see parental education as a parental socialization element, which forms ethnic German identity which then forms the decision of how much German education an individual is willing to gain. We see the effect of parental education on education in Germany is an effect which happens via the German ethnic identity an individual develops. The same is true for the time spent in Germany. It does not directly affect educational decisions but via the degree of integration - the degree of German ethnic identity - an individual has actually reached.

Table 4.3. shows that all three nationality groups of the second generation - Turkish, Central European and EU15 identify much more with Germany than EU15 immigrants. Second generation immigrants identify by about 2 units more than immigrants. This is a considerable amount since the scale of identity ranges from -1.2 to 3.5. Another interesting observation is that the coefficients for the different nationality groups among the second generation do not differ much. This could imply that difference in identification with Germany across nationality groups dies out over generations. The first generation still seems to differ across nationalities in their identity levels: Turkish citizens identify by about 1.3 units less with Germany than EU15 immigrants but there is only a small insignificant difference between central European and EU15 immigrants. Women seem to identify slightly more with Germany. Education in 1999 has a positive and significant impact on identity. Being still in education or having a high educational level does not have a significant effect on German identity. Low educational attainment in 1999 has a significantly negative impact on identity. Mother's educational attainment[11] seems to be more important than father's education. It has a positive impact on German identity if the mother has obtained a higher degree than

[11]The base categories for the parental education variables are low educational attainment level of the father and of the mother respectively.

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

an upper secondary degree. Conversely it has a negative impact on German identity if the mother has not obtained any degree. We also control for the length of stay in Germany since to test the adaptation hypothesis. The results show that the longer the immigrant or second generation immigrant has been living in Germany since 1999, the stronger he or she identifies with Germany.

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

β^D,α^D	(1)	(2)	(3)
Intercept	0.12	-0.11	0.13
	(-1.01,1.22)	(-1.31,1.06)	(-1.08,1.37)
Age	-0.01	0.01	0.01
	(-0.0.4,0.03)	(-0.02,0.04)	(-0.03,0.04)
Gender	-0.28	-0.32	-0.34
	(-0.54,-0.03)	(-0.58,-0.05)	(-0.60,-0.07)
Second generation	0.38	0.04	
	(0.11,0.66)	(-0.27,0.34)	
Central European immigrant			0.43
			(-0.04,0.93)
Turkish immigrant			-0.32
			(-0.74,0.12)
Turkish second generation			-0.29
			(-0.84,0.25)
Central European second generation			-0.28
			(-0.93,0.38)
EU15 second generation			0.03
			(-0.39,0.45)
Identity		0.35**	0.30**
		(0.23,0.46)	(0.18,0.43)

**p< .05; Bayesian confidence interval in parentheses

Table 4.1: Estimates of the Education Equation: $D^* = \alpha\theta + \beta X + \varepsilon_D$, estimated by MCMC

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

α^M	(1)	(2)
Spoken german	1.72	1.78
	(1.10,2.54)	(1.15,2.51)
Written german	1.03	1.03
	(0.75,1.33)	(0.76,1.33)
Language used	0.87	0.83
	(0.69,1.04)	(0.67,1.01)
Feel german	0.44	0.43
	(0.35,0.544)	(0.33,0.53)
Newspaper	0.86	0.83
	(0.70,1.03)	(0.67,0.99)
Food	0.35	0.33
	(0.26,0.44)	(0.25,0.42)
Friends german	0.22	0.22
	(0.14,0.30)	(0.14,0.29)

Table 4.2: Estimates of the Psychometric Question Equations: $M^* = \alpha\theta + \varepsilon_M$

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

γ	(1)	(2)
Age	-0.00	-0.00
	(-0.02,0.019)	(-0.02,0.02)
Gender	0.06	0.10
	(-0.20,0.34)	(-0.18,0.37)
Turkish immigrant	-1.30	-1.27
	(-1.89,-0.69)	(-1.88,-0.67)
Central European immigrant	0.04	0.02
	(-0.60,0.65)	(-0.62,0.64)
Turkish second generation	2.31	2.31
	(1.66,2.95)	(1.65,2.97)
Central European second generation	2.27	2.36
	(1.51,3.00)	(1.59,3.12)
EU15 second generation	2.25	2.28
	(1.42,3.09)	(1.43,3.097)
German second generation	2.07	2.05
	(1.24,2.89)	(1.22,2.86)
Low education	-0.38	-0.36
	(-0.68,-0.067)	(-0.67,-0.053)
High education	0.40	0.44
	(-0.12,0.94)	(-0.093,0.97)
Father medium education level	0.01	0.00
	(-0.39,0.38)	(-0.45,0.45)
Mother medium education level	0.09	0.10
	(-0.26,0.50)	(-0.35,0.55)
Father no schooling	0.14	0.16
	(-0.74,1.00)	(-0.73,1.04)
Mother no schooling	-1.64	-1.66
	(-2.70,-0.53)	(-2.72,-0.55)
Father highly education	0.21	0.22
	(-0.20,0.58)	(-0.19,0.62)
Mother highly educated	0.50	0.51
	(0.10,0.89)	(0.13,0.90)
In education in 1999	0.45	0.39
	(-0.02,0.89)	(-0.08,0.85)
Time stayed in Germany	0.09	0.09
	(0.056,0.12)	(0.06,0.12)

Table 4.3: Estimates of the Psychometric Question Equations: $\theta = \gamma W + \varepsilon_{theta}$

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

The results of the model show that German identity at a given point in time and with a given level of educational attainment at that moment is a significant determinant for future educational attainment. German identity can compensate for the immigrant generational difference in educational levels. Immigrants seem to have a double disadvantage compared to the second generation: as shown in table 4.1 immigrants have a lower probability of having a tertiary degree and immigrants identify significantly less with Germany than the second generation, which also reduces their probability of having a tertiary degree. An interesting result for policy makers is also the importance of mother's education. This finding supports the policy to target immigrant mothers in order to increase educational attainment of immigrant children.

4.5 Conclusion

Integrating immigrants into the German labor market can be beneficial both for the immigrant and for Germany : Germany needs high skilled labor to keep up the levels of economic growth and immigrants need jobs. German schools have the important task to make adolescents with a migration background fit for the labor market of their host country and policy makers should increase educational attainment levels of immigrants and their children. The key for integrating immigrants and their children into the German labor market is a successful education. For education to be successful firstly the immigrant needs to be ready to learn and to provide the effort to go into higher tracks of the German education system, and secondly the German government needs to provide an educational system compatible for immigrants - for individuals with a different or double cultural background.

We study the role that a day-to-day life German identity - measured by examining practical integration measures - plays for educational attainment of immigrants and their children. By employing a continuous latent factor as a measure for German identity, we use a more precise and efficient measure than by using simply the questions in the questionnaire. We use more information and allow for a continuum of identity outcomes. Our identity measure is assumed to be endogenous. This methodology addresses the important problem of endogeneity of German identity in an educational outcome equation and gives insight on the determinants of German identity.

The chapter finds that a strong German identity measured at a specific point in time increases the probability of having a tertiary degree in the future and can compensate the disadvantage that immigrants have in terms of tertiary education compared to the second generation. Our findings hold when controlling for the educational attainment level at the time German identity is measured and for other determinants of identity. We find that

second generation immigrants have a stronger German identity, no matter of which ethnic background. They are more integrated in terms of friends, visible cultural practices, language issues and even in terms of their ethnic self-identification than their parents. Differences in German identity across ethnic groups are no longer present for the second generation. We also find that mother's education is an important determinant of German identity.

This chapter shows that one way to increase educational attainment of immigrants and their children can be to increase their identification with Germany for example by increasing the contacts between immigrants and their children with natives, by increasing language classes and by interesting immigrants in the German culture and media. The paper also shows that a way to increase German identity of immigrants and their children is to target their mothers since an educated mother increases German identity on average.

The integration of immigrants and their children is multi-dimensional as shown in chapter one. This chapter aims to link an economically important dimension - educational attainment - with a closely related sociological dimension - identity - by using an econometric methodology that enables this multidisciplinary approach and addresses its most common problems of measurement error and endogeneity.

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

4.6 Appendix A:

4.6.1 Estimation : The Gibbs Sampler

Bayesian MCMC (Markov Chain Monte Carlo) methods allow to simulate from a posterior distribution function,which is considered - in line with Bayesian statistics -to be proportional to prior distributions and the likelihood function of the model. In our case this methodology is useful since the likelihood function contains an integral over the latent variables. We follow a methodology based upon Carneiro, Hansen, Heckman (2003) and Heckman, Stixrud and Urzua (2006) and employ the Gibbs sampler, one of the most popular MCMC algorithms. Even though the method originates from Bayesian statistics, the Bayesian ideology need not be adopted and the method can simply be used for computational convenience as Carneiro, Hansen and Heckman (2003) and Heckman, Stixrud and Urzua (2006) note.

The posterior joint distribution for individual i of the model can be written as

$$\begin{aligned} & f(\beta,\alpha,\gamma,\theta_i,Y_i^*,M_i^*,c|Y_i,M_i,X_i,W_i) \\ \propto\ & f(\beta)f(\alpha)f(\gamma)f(c)f(M_i,Y_i,Y_i^*,M_i^*,\theta_i|X_i,W_i,\beta,\alpha,\gamma,c) \end{aligned}$$

The likelihood function can be simplified as

$$\begin{aligned} & f(M_i,Y_i,Y_i^*,M_i^*,\theta_i|X_i,W_i,\beta,\alpha,\gamma,c) \\ = \ & f(Y_i^*,M_i^*,\theta_i|X_i,W_i,\beta,\alpha,\gamma,c)f(M_i,Y_i|Y_i^*,M_i^*,\theta_i,X_i,W_i,\beta,\alpha,\gamma,c) \\ = \ & f(Y_i^*,M_i^*,\theta_i|X_i,W_i,\beta,\alpha,\gamma,c)f(M_i,Y_i|c) \end{aligned}$$

The likelihood functions of M_i and Y_i written separately are

$$\begin{aligned} f(Y_i^*,\theta_i|\alpha,\beta,\gamma,c,Y_i,X_i,W_i)\{\sum_{k_Y=1}^{K_Y} 1(Y_i^* &= k_Y)1(c_{k_Y-1}<Y_i^*<c_{k_Y})\} \\ f(M_i^*,\theta_i|\alpha,c,M_i,W_i)\{\sum_{k_M=1}^{K_M} 1(M_i^* &= k_M)1(c_{k_M-1}<M_i^*<c_{k_M})\} \end{aligned}$$

When factoring out the latent factor θ_i we can write

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

$$
\begin{aligned}
f(M_i^*|\alpha, c, M_i, W_i) &= \int_\theta f(M_i^*|\alpha, c, \theta_i, M_i) f(\theta_i|W_i) d\theta \\
f(Y_i^*|\alpha, c, Y_i, X_i, W_i) &= \int_\theta f(Y_i^*|\alpha, c, \theta_i, Y_i, X_i) f(\theta_i|W_i) d\theta
\end{aligned}
$$

We estimate the joint posterior distribution of the model by a Gibbs sampler. In the following we derive the full conditional distributions of the model.

4.6.1.1 The Posterior Conditional Distribution of the Latent Underlying Variables

Albert and Chib (1993) propose a data augmentation procedure to sample latent underlying variables in a threshold model. It follows from his work, that the full conditional for the latent underlying variable of the polytomous responses of the economic outcome variable and the psychometric measures are

$$
\begin{aligned}
& f(Y_i^*|\alpha, \beta, \theta_i, c, Y_i, X_i) \\
\propto & \prod_{i=1}^{N} f(Y_i^*|\beta^Y X_i + \alpha^Y \theta_i, 1) \{ \sum_{k_Y=1}^{1} 1(Y_i^* = k_Y) 1(c_{k_Y - 1} < Y_i^* < c_{k_Y}) \} \\
& f(M_i^*|\alpha, \beta, \theta_i, c, M_i) \\
\propto & \prod_{i=1}^{N} f(M_i^*|\alpha^M \theta_i, 1) \{ \sum_{k_M=1}^{K_M} 1(M_i^* = k_M) 1(c_{k_M - 1} < M_i^* < c_{k_M}) \}
\end{aligned}
$$

where $V(Y_i^*)$ is normalized to 1. The latent underlying variables are distributed as the following truncated normal distributions

$$
\begin{aligned}
Y_i^*|\alpha, \beta, \theta_i, c, Y_i, X_i &\sim TN_{(c_{k_Y - 1}, c_{k_Y})}(\beta^Y X_i + \alpha^Y \theta_i, 1) \\
M_i^*|\alpha, \beta, \theta_i, c, M_i &\sim TN_{(c_{k_M - 1}, c_{k_M})}(\alpha^M \theta_i, 1)
\end{aligned}
$$

4.6.1.2 The Posterior Conditional Distribution of the Factor Loadings

The full conditional for the factor loadings for Y can be written as

$$
f(\alpha^Y|\beta, \theta_i, Y_i, X_i, Y_i^*) \propto f(\alpha^Y) \prod_{i=1}^{N} f(Y_i^*|\beta^Y X_i^Y + \alpha^Y \theta_i, 1)
$$

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

where we choose a normal prior $f(\alpha^Y) = N(0,1)$.If we rewrite the equation for Y^* as

$$Y_i^* - \beta^Y X_i^Y = \alpha^Y \theta_i + \varepsilon_i^Y$$

we can treat it as a normal regression model and derive

$$\alpha^Y|\beta, \theta_i, Y_i, X_i, Y_i^* \sim N\left[(\theta_i'\theta_i + 1)^{-1}\theta_i'(Y_i^* - \beta^Y X_i^Y), (\theta_i'\theta_i + 1)^{-1}\right]$$

Similarly for M_i with prior $f(\alpha^M) = N(0,1)$ we can write

$$\alpha^M|\theta_i, M_i, M_i^* \sim N\left[(\theta_i'\theta_i + 1)^{-1}\theta_i'(M_i^*), (\theta_i'\theta_i + 1)^{-1}\right]$$

4.6.1.3 The Posterior Conditional Distribution of the Direct Coefficients

Similarly to the procedure for the factor loadings, we can write the model as

$$Y_i^* - \alpha^Y \theta_i = \beta^Y X_i^Y + \varepsilon_i^Y$$

With prior $f(\beta^Y) = N(0,1)$ we can write the full conditional for the direct coefficients as

$$\beta^Y|\alpha^Y, \theta, c, Y, M, D, X, Y^*, D^*, M^* \sim N\left[(X_i'X_i + 1)^{-1}X_i'(Y_i^* - \alpha^Y\theta_i^Y), (X_i'X_i + 1)^{-1}\right]$$

4.6.1.4 The Posterior Conditional Distribution of the Cutpoints

We assume a uniform prior for the cutpoints and can write for the full conditionals for the polytomous responses

$$\begin{aligned} c^M|\alpha^M, \theta_i, M_i, M_i^* &\sim unif\begin{bmatrix} \max\{\max\{M_i^* : M_i = k_M\}, c_{M-1}\}, \\ \min\{\min\{M_i^* : M_i = k_{M+1}\}, c_{M+1}\} \end{bmatrix} \\ c^Y|\alpha^Y, \beta^Y, \theta_i, Y_i, X_i, Y_i^* &\sim unif\begin{bmatrix} \max\{\max\{Y_i^* : Y_i = k_Y\}, c_{Y-1}\}, \\ \min\{\min\{Y_i^* : Y_i = k_{Y+1}\}, c_{Y+1}\} \end{bmatrix} \end{aligned}$$

4.6.1.5 The Posterior Conditional Distribution of the Latent Factors

Similarly as for the procedure for coefficients and factor loadings, we can rewrite the model as

$$\begin{aligned} Y_i^* - \beta^Y X_i^Y &= \alpha^Y \theta_i + \varepsilon_i^Y \\ M_i^* - \beta^M X_i^M &= \alpha^M \theta_i + \varepsilon_i^M \end{aligned}$$

and treat it as a normal regression model,where θ_i is the parameter to be estimated. We can then derive the full conditional for the latent factor as:

$$\begin{aligned} & f(\theta|\beta, \alpha, c, X, W, Y^*, D^*, M^*) \\ \propto & \prod_{i=1}^{N} f(Y_i^*|\beta^Y X_i^Y + \alpha^Y \theta_i, 1) f(M_i^*|\alpha^M \theta_i, 1) \end{aligned}$$

$$\begin{aligned} & \theta_i|\beta, \alpha, \gamma, c, X_i, Y_i^*, M_i^*, W_i^* \\ \sim & N \left[\begin{array}{c} \gamma W_i + (\alpha^{Y\prime}\alpha^Y + \alpha^{M\prime}\alpha^M + 1)^{-1} \\ (\alpha^{Y\prime}(Y_i^* - \beta^Y X_i^Y - \alpha^{Y\prime}\gamma W_i) + \alpha^{M\prime}(M_i^* - \alpha^{M\prime}\gamma W_i)), \\ I - \alpha^{Y\prime}(\alpha^{Y\prime}\alpha^Y + \alpha^{M\prime}\alpha^M + 1)^{-1}\alpha^Y \\ -\alpha^{M\prime}(\alpha^{Y\prime}\alpha^Y + \alpha^{M\prime}\alpha^M + 1)^{-1}\alpha^M \end{array} \right] \end{aligned}$$

4.6.1.6 The Posterior Conditional Distribution of the Indirect Coefficients

The posterior we sample from can be written as

$$\begin{aligned} & f(\gamma|\theta, W) \\ \propto & f(\gamma) f(\theta|\gamma, W) \end{aligned}$$

The model for the latent variable is

$$\theta = \gamma W + \varepsilon^\theta$$

We assume a diffuse prior for the coefficient γ. Similar to the procedures above we get:

$$f(\gamma|\theta, W) \sim N((W'W)^{-1}W'\theta), (W'W)^{-1})$$

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

4.7 Appendix B: Descriptive Tables

Educational Attainment in 1999,Second Generation aged 17-20					
	In School	Low	Medium	High	Total
25th identity percentile	1	1	0	0	2
in percent	50	50	0	0	100
50th identity percentile	3	0	0	0	3
in percent	100	0	0	0	100
75th identity percentile	12	4	3	0	19
in percent	63	21	16	0	100
100th identity percentile	14	2	2	0	18
in percent	78	11	11	0	100
total	30	7	5	0	42
in percent	71	17	12	0	100

Educational Attainment in 2007,Second Generation aged 17-20					
	In School	Low	Medium	High	Total
25th identity percentile	0	2	0	0	2
in percent	0	100	0	0	100
50th identity percentile	0	1	2	0	3
in percent	0	33	67	0	100
75th identity percentile	0	4	12	3	19
in percent	0	21	63	16	100
100th identity percentile	0	4	12	2	18
in percent	0	22	67	11	100
total	0	11	26	5	42
in percent	0	26	62	12	100

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

Educational Attainment in 1999,Immigrants aged 17-20					
	In School	Low	Medium	High	Total
25th identity percentile	0	0	0	0	0
in percent	0	0	0	0	100
50th identity percentile	1	0	0	0	1
in percent	100	0	0	0	100
75th identity percentile	1	0	0	0	1
in percent	100	0	0	0	100
100th identity percentile	1	0	0	0	100
in percent	100	0	0	0	100
total	3	0	0	0	3
in percent	0	0	0	0	100

Educational Attainment in 2007,Immigrants aged 17-20					
	In School	Low	Medium	High	Total
25th identity percentile	0	0	0	0	0
in percent	0	0	0	0	0
50th identity percentile	0	0	1	0	1
in percent	0	0	100	0	100
75th identity percentile	0	1	0	0	1
in percent	0	100	0	0	100
100th identity percentile	0	1	0	0	100
in percent	0	100	0	0	100
total	0	2	1	0	3
in percent	0	67	33	0	100

Educational Attainment in 1999,Natives aged 17-20						
	In School	Low	Medium	High	Total	Missing
total	218	59	37	1	8	323
in percent	67	18	11	0	2	100

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

Educational Attainment in 2007,Natives aged 17-20						
	In School	Low	Medium	High	Total	Missing
total	0	33	234	56	0	323
in percent	0	10	72	17	0	100

Educational Attainment in 1999,Second Generation aged 21-26						
	In School	Low	Medium	High	Mising	Total
25th identity percentile	0	4	1	0	0	5
in percent	0	80	20	0	0	100
50th identity percentile	2	9	15	0	2	28
in percent	7	32	54	0	7	100
75th identity percentile	3	5	9	0	2	19
in percent	16	26	47	0	11	100
100th identity percentile	2	3	12	0	1	18
in percent	11	17	67	0	6	100
total	7	21	37	0	5	70
in percent	10	30	53	0	7	100

Educational Attainment in 2007,Second Generation aged 21-26						
	In School	Low	Medium	High	Mising	Total
25th identity percentile	0	3	2	0	0	5
in percent	0	60	40	0	0	100
50th identity percentile	0	7	19	0	2	28
in percent	0	25	68	7	0	100
75th identity percentile	0	3	11	5	0	19
in percent	0	16	58	26	0	100
100th identity percentile	0	2	9	7	0	18
in percent	0	11	50	39	0	100
total	0	15	41	14	0	70
in percent	0	21	59	20	0	100

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

Educational Attainment in 1999,Immigrants aged 21-26						
	In School	Low	Medium	High	Mising	Total
25th identity percentile	0	3	5	1	5	14
in percent	0	21	36	7	36	100
50th identity percentile	1	3	8	0	1	13
in percent	8	23	62	0	8	100
75th identity percentile	0	0	3	1	0	4
in percent	0	0	75	25	0	100
100th identity percentile	1	1	2	1	0	5
in percent	20	20	40	20	0	100
total	2	7	18	2	7	36
in percent	6	19	50	6	19	100

Educational Attainment in 2007,Immigrants aged 21-26						
	In School	Low	Medium	High	Mising	Total
25th identity percentile	0	8	5	1	0	14
in percent	0	57	36	7	0	100
50th identity percentile	0	2	10	1	0	13
in percent	0	15	77	8	0	100
75th identity percentile	0	1	2	1	0	4
in percent	0	25	50	25	0	100
100th identity percentile	0	0	3	2	0	5
in percent	0	0	60	40	0	100
total	0	11	20	5	0	36
in percent	0	31	56	14	0	100

Educational Attainment in 1999,Natives aged 21-26						
	In School	Low	Medium	High	Mising	Total
total	32	74	330	46	35	517
in percent	6	14	64	9	7	100

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

Educational Attainment in 2007,Natives aged 21-26						
	In School	Low	Medium	High	Mising	Total
total	0	63	282	172	0	517
in percent	0	12	55	33	0	100

Educational Attainment in 1999,Second Generation aged 27-32						
	In School	Low	Medium	High	Mising	Total
25th identity percentile	0	9	4	1	2	16
in percent	0	56	25	6	13	100
50th identity percentile	0	6	8	1	2	17
in percent	0	35	47	6	12	100
75th identity percentile	0	5	8	2	0	15
in percent	0	33	53	13	0	100
100th identity percentile	0	4	13	9	1	27
in percent	0	15	48	33	4	100
total	0	24	33	13	5	75
in percent	0	32	44	17	7	100

Educational Attainment in 2007,Second Generation aged 27-32						
	In School	Low	Medium	High	Mising	Total
25th identity percentile	0	9	5	2	0	16
in percent	0	56	31	13	0	100
50th identity percentile	0	7	7	3	0	17
in percent	0	41	41	18	0	100
75th identity percentile	0	4	8	3	0	15
in percent	0	27	53	20	0	100
100th identity percentile	0	1	15	11	0	27
in percent	0	4	56	41	0	100
total	0	21	35	19	0	75
in percent	0	28	47	25	0	100

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

Educational Attainment in 1999,Immigrants aged 27-32						
	In School	Low	Medium	High	Mising	Total
25th identity percentile	0	23	12	0	5	40
in percent	0	58	30	0	13	100
50th identity percentile	0	7	3	2	3	15
in percent	0	47	20	13	20	100
75th identity percentile	0	7	8	4	0	19
in percent	0	37	42	21	0	100
100th identity percentile	0	1	6	0	0	7
in percent	0	14	86	0	0	100
total	0	38	29	6	8	81
in percent	0	47	36	7	10	100

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

Educational Attainment in 2007,Immigrants aged 27-32						
	In School	Low	Medium	High	Mising	Total
25th identity percentile	0	24	15	1	0	40
in percent	0	60	38	3	0	100
50th identity percentile	0	6	6	3	0	15
in percent	0	40	40	20	0	100
75th identity percentile	0	4	10	5	0	19
in percent	0	21	53	26	0	100
100th identity percentile	0	1	5	1	0	7
in percent	0	14	71	14	0	100
total	0	35	36	10	0	81
in percent	0	43	44	12	0	100

Educational Attainment in 1999,Natives aged 27-32						
	In School	Low	Medium	High	Mising	Total
total	8	132	482	197	67	886
in percent	1	15	54	22	8	100

Educational Attainment in 2007,Natives aged 27-32						
	In School	Low	Medium	High	Mising	Total
total	0	96	495	295	0	886
in percent	0	11	56	33	0	100

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

BIBLIOGRAPHY

[1] Akerlof, G. & Kranton, R. (2000) : Economics and Identity, *Quarterly Journal of Economics, 115(3), 715-753, MIT Press.*

[2] Akerlof, G. & Kranton, R. (2002) : Identity and Schooling - Some Lessons from the Economics of Education, *Journal of Economic Literature, 40 (4), 1167-1201.*

[3] Akerlof, G. & Kranton, R. (2010): Identity Economics: How our identities shape our work, wages and well-being. *Princeton University Press.*

[4] Albert, J.H. & Chib, S. (1993) : Bayesian Analysis of Binary and Polychotomous Response Data, *Journal of the American Statistical Association 88 (422).*

[5] Becker, G. (1964) : Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education. Chicago, University of Chicago Press.

[6] Berry, J.W. (1980) : Acculturation as Varieties of Adaptation in A.M. Padilla (Ed.), *Acculturation : Theory, Models and Some New Findings,* (pp. 9-25). Boulder, CO: Westview.

[7] Bonin, H.; Constant, A.; Tatsiramos, & Zimmermann, K.F. (2006) : Ethnic Persistence, Assimilation and Risk Proclivity, *IZA Discussion Papers, Institute for the Study of Labour (IZA).*

[8] Borjas, G. (1992) : Ethnic Capital and Intergenerational Mobility, *Quarterly Journal of Economics.*

Thum, Anna-Elisabeth (2011), Psychological Factors and Labour Market Outcomes: The case of immigrants and their children in Germany
European University Institute DOI: 10.2870/24737

[9] Cameron, S. & Heckman, J. (1999) : The Dynamics of Educational Attainment for Blacks, Hispanics and Whites, *NBER working paper.*

[10] Carneiro, P.; Hansen, K. & .Heckman, J. (2003) : Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice, *International Economic Review.*

[11] Chiswick, B.R. (1978) : The Effect of Americanization on Earnings of Foreign-born Young Men, *Journal of Political Economy 86(5), 897-921.*

[12] Chiswick, B.R. (1988) : Differences in Education and Earnings across Racial and Ethnic Groups : Tastes, Discrimination, and Investments in Child Quality, *Quarterly Journal of Economics.*

[13] Chiswick, B.R. & Miller, P.W. (1994) : The Determinants of Post-Immigration Investments in Education, *Economics of Education Review.*

[14] Chiswick, B.R. & DebBurman, N. (2003) : Educational Attainment : Analysis by Immigrant Generation, *Economics of Education Review.*

[15] Constant, A.; Gataullina, L. & Zimmermann, K.F. (2006) Ethnosizing Immigrants, *IZA discussion paper, Institute for the Study of Labor (IZA).*

[16] Constant, A.; Gataullina, L. & Zimmermann, K.F. (2006) : Gender, Ethnic Identity and Work, *IZA discussion paper, Institute for the Study of Labor (IZA).*

[17] Constant, A. & Zimmermann, K.F. (2006) : Ethnic Self-identification of First-Generation Immigrants, *IZA discussion paper, Institute for the Study of Labor (IZA).*

[18] Constant, A. & Zimmermann, K.F. (2007) : Measuring Ethnic Identity and its Impact on Economic Behavior, *IZA discussion paper, Institute for the Study of Labor (IZA).*

[19] Duleep, H. & Regrets, M. (1999) : Immigrants and Human Capital Investment, *American Economic Review.*

[20] Epstein, Z. & Gang, I. (2010): Migration and Culture, *IZA discussion paper 5123, Institute for the Study of Labor (IZA).*

[21] Erikson, E.H. (1950) : Childhood and Society. New York : Norton.

[22] Erikson, E.H. (1959) : Identity and the Life Cycle; selected papers. New York : International Universities Press.

[23] European Commission (2010) : Europe 2020 - A strategy for smart, sustainable and inclusive growth, Communication from the Commission, Brussels 2010.

[24] Fahrmeir, L. & Raach, A. (2006): A Bayesian semiparametric latent variable model for mixed responses, *Psychometrika.*

[25] Fertig, M. & Schmidt, C. (2001) : First and Second-Generation Migrants in Germany – What do we know and what do People Think?, *IZA discussion paper, Institute for the Study of Labor (IZA).*

[26] Fuchs, M. & Sixt, M. (2008): Die Bildungschancen von Aussiedlerkindern, *SOEP papers on Multidisciplinary Panel Data Research, 105.*

[27] Gang, I. & Zimmermann, K.F. (1999) : Is Child like Parent? Educational Attainment and Ethnic Origin, *Journal of Human Resources.*

[28] Gataullina, L. & Zimmermann, K.F. (2006), Human Capital and Ethnic Selfidentification of Migrants, *IZA Discussion Papers, Institute for the Study of Labor (IZA).*

[29] Heckman, J.; Stixrud, J. & Urzua, S. (2006) : The Effects of Cognitive and Non-cognitive Abilities on Labor Market Outcomes and Social Behavior, *Journal of Labor Economics*

[30] Matzkin, R. (2003) : Unobservable Instruments, *mimeo*, Northwestern University.

[31] Matzkin, R. (2007) : Nonparametric Identification, *Handbook of Econometrics Vol 6B.*

[32] Neckby, Roedlin (2007): Acculturation Identity and Educational Attainment, *IZA Discussion Papers 2826, Institute for the Study of Labor (IZA).*

[33] Phinney (1992) : The Multigroup Ethnic Identity Measure : A new Scale for use with Diverse Ethnic Groups , *Journal of Adolescent Research.*

[34] Rabe-Hesketh, S. & Skondral, A. (2004) : Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models. Boca Raton, FL: Chapman & Hall/CRC.

[35] Riphahn (2001) : Dissimilation? The Educational Attainment of Second-Generation Immigrants, *CEPR discussion paper.*

[36] Worrell (2006) : Multigroup Ethnic Identity Measure Scores in a Sample of Adolescents from Zimbabwe*, Identity : An International Journal of Theory and Research.*